Proceedings of the International Computer Science and Technology Conference 2008

Edited by
John Bugado, Mohammad Amin, Pradip Peter Dey,
Chuck Brown, and Arun Datta
Preface

Welcome to the International Computer Science and Technology Conference 2008 (ICSTC). This conference is intended to bring together researchers and practitioners from industry, academia and government to advance the state of the art in computer science, software engineering and technology and to encourage wider collaboration between academics and industry. The Conference is held for the computing community and hosted by National University. The large number of submitted papers is a clear indication of enthusiastic cooperation and response from the community. Out of the 183 submissions, 55 papers were initially accepted based on recommendations from our reviewers. However, many papers were subsequently withdrawn by mostly international authors due to some serious reasons including, visa and international situations. The program committee made their best efforts to accommodate all submissions with academic merit and scholarship.

National and internationally recognized speakers will present their research contributions in the field. In addition, a number of panelists will attend for the following six panel sessions: 1) Multi-model Multi-strategy Teaching/Learning in Science, Engineering and Technology, 2) Database Systems for Health Informatics, 3) Regulatory Compliance, 4) Strategic Management, 5) Green Computing, and 6) Physical Data Protection Panel. There will be two special presentations: 1) Virtualization by Mr. Darel Ison, Director of Technical Services at GTC Systems and 2) Business Process Management by Mr. Neal Fischer is the Founder of Hershey Technologies.

Finally, we would like to acknowledge the support and cooperation of all the authors and reviewers of ICSTC-08. We are thankful to our distinguished keynotes speakers: Dr. Dana Gibson, Dr. Peter Arzberger, Chair of the Pacific Rim Application and Grid Middleware Assembly, Dr. Subhas C. Misra, Director of Web World Network, Mr. Norbert J. Kubilus, COO of DataLEAD Communications and the Network Services Division of National Data Corporation, and Todd Walter, CTO, Teradata Research & Development. Our special thanks to Dr. J. Lee, Chancellor of National University System, Dr. Tom Green, VPAA of National University and Dr. Howard Evans, Dean of School of Engineering and Technology for their valuable suggestions and supports.

Conference Chair: John Bugado
Program Chairs: Chuck Brown, Pradip Peter Dey, Mohammad Amin, and Arun Datta
Event Management Chair: Albert Cruz

International Computer Science and Technology Conference (WWW.ICSTC.ORG)
San Diego, California, USA
April 1, 2008
<table>
<thead>
<tr>
<th>1. Structuring Meta-search Research by Design Patterns</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jürgen Dorn and Tabbasum Naz</td>
<td></td>
</tr>
<tr>
<td>2. APRIORI_PROCREATION- TOOL for Optimizing WEB PERFORMANCE</td>
<td>17</td>
</tr>
<tr>
<td>Aruna Jain</td>
<td></td>
</tr>
<tr>
<td>3. Formal Privacy Ontology, a Definition</td>
<td>28</td>
</tr>
<tr>
<td>Wael A. Hassan</td>
<td></td>
</tr>
<tr>
<td>4. A Method for Optimizing the Size of Wireless Mobile Messages</td>
<td>40</td>
</tr>
<tr>
<td>P P Abdul Haleem and M P Sebastian</td>
<td></td>
</tr>
<tr>
<td>5. Test Case Generation for White-Box Unit Testing</td>
<td>48</td>
</tr>
<tr>
<td>Samira Hussain and Salam N. Salloum</td>
<td></td>
</tr>
<tr>
<td>6. Integrating Generative and Aspect-Oriented Technologies for Framework Instantiation</td>
<td>59</td>
</tr>
<tr>
<td>R. P. Mahapatra and Dr. Jayaram Pradhan</td>
<td></td>
</tr>
<tr>
<td>7. Semi-Automatic Generation of Transformation Rules in Model Driven Engineering: The Challenge and First Steps</td>
<td>72</td>
</tr>
<tr>
<td>Slimane Hammoudi and Samyr Vale</td>
<td></td>
</tr>
<tr>
<td>8. Use of Simulation in Systems Engineering</td>
<td>80</td>
</tr>
<tr>
<td>Hassan Badkoobehi</td>
<td></td>
</tr>
<tr>
<td>9. Content Delivery Network: A Dominant Technology for Internet</td>
<td>85</td>
</tr>
<tr>
<td>Tulika Srivastava and Mudita Srivastava</td>
<td></td>
</tr>
<tr>
<td>10. Secure Aroma Transfer</td>
<td>91</td>
</tr>
<tr>
<td>Vibhore Kumar Jain, Subha Joshi and Chaman Kumar Bansal</td>
<td></td>
</tr>
<tr>
<td>11. RESOURCE DESCRIPTION FRAMEWORK : A SMALL CASE STUDY ON EDUCATIONAL DOMAIN</td>
<td>98</td>
</tr>
<tr>
<td>V. Kiran Kumar</td>
<td></td>
</tr>
<tr>
<td>12. Cooperative Intrusion Detection for Detecting Novel Attacks Using Real Time Data Mining Approach</td>
<td>104</td>
</tr>
<tr>
<td>R. Pushparaj</td>
<td></td>
</tr>
<tr>
<td>13. Data Privacy and Technical Solutions, Iqbal Khan,</td>
<td>120</td>
</tr>
<tr>
<td>14. Application of Cognitive Styles and Inference to Project Management,</td>
<td>121</td>
</tr>
<tr>
<td>David J. Sudbeck</td>
<td></td>
</tr>
<tr>
<td>15. Early Muslim Scholars and Binary Mathematics</td>
<td>128</td>
</tr>
</tbody>
</table>
Chetankumar Patel and Muthu Ramachandran  
133

17. Analysis of Near-infrared Phase Effect on Biometric Iris Data Demonstrates Viability of Iris-scan Biometric Device for Identity Authentication in Public Venues  
Brady R. Stevenson and Gordon W. Romney  
142

18. Challenges in Vendor Collaboration in A Mix Mode Offshore Centre - Coexistence of Captive And Vendor Offshore  
Satyen Vats and Rathi Dasgupta  
143

19. COMPARATIVE STUDY AND ANALYSIS OF FLANK WEAR OF SINGLE POINT COATED AND UNCOATED TURNING TOOL INSERTS USING ACOUSTIC EMISSION SIGNAL PROCESSING AND NEURAL NETWORK TECHNIQUES  
N.PRASANNA, R. SRINIDHI, M.SUKUMAR, and C.S.VENKATESHA  
149

20. Multi-model Multi-strategy Teaching/Learning in Science, Engineering and Technology  
Pradip Peter Dey, Mohammad Amin, Michelle Bright, Arun Datta, Shakil Akhtar, Albert Cruz, Mudasser F. Wyne, Patrick Kennedy, Patrick Olson, Rell Snyder, Alireza Farahani, Samuel Afuwape, Hassan Badkoobehi and Amber Lo  
158

21. Exploration Strategy Based on Terrain Coverage for Multiple Robots Using Genetic Algorithm  
K.S. Senthilkumar and K. K. Bharadwaj  
163

22. Capitalising Software-Engineering Tacit Knowledge  
Abdulmajid Hissen Mohamed  
170

23. On Simplification of Grid Scheduling and Resource Discovery Problem A Survey and Proposal  
Waqas Ahmed, Shaleeza Sohail and Aasia Khanum  
175

24. Joint Modeling and Optimization of Business Processes and Information Technologies  
Igor Safonov and Vadim Safonov  
185

25. Prediction Based Dynamic Load Balancing Techniques in Heterogeneous Clusters  
Pushpendra Kumar Chandra and Bibhudatta Sahoo  
189

26. QUERY PROCESSING IN MAIN MEMORY OODB SYSTEMS  
193
27. Rapid Prototyping Data Integration
   Barbara Lauridsen

28. Relational Database Loader for Implementing Advanced Queries in
    Protein Informatics
   Sharen A. Eskelsen, Scott S. Kelman, and Carl S. Ewig

29. Categorization of Functional Dependencies for a Minimal Cover
    Ajeet A. Chikkamanur and Shivanand M. Handigund

30. CHOMS: A Portal for Obesity Management
    Radha Nandkumar, Arun Datta, Jacqueline Caesar, Weimo Zhu,
    Peter Bajcsy, Jeff Sunderland, Victoria Jackson, Jill Sproat, and Heidi Krahling

31. Why is C4I Software Hard to Develop?
    Lee Whitt

32. Waveform Distortion Estimation during Multiple Power Quality Events
    Amritpal Singh and Manmohan Singh

33. How to secure e-mail and instant messages
    Bari Siddique, Rozina Siddique, and Mohammad Amin

34. Database Panel Discussions: Database Systems for Health Informatics
    Moderator: Jacqueline Caesar

35. Integrating Multiple Data Sources Into One Analytical Environment
    Imad Birouty

36. Adaptive Neuro-Fuzzy Inference System based Inverse Kinematics Solution
    of 3DOF Planar Robot
    Srinivasan Alavandar and M.J.Nigam

37. Intelligent Cash Management System for an ATM network
    Rimvydas Simutis, Darius Dilijonas, and Lidija Bastina

38. RIN Evaluation with dispersion compensation for Optimum Performance
    of Optical Transmission System
    Kamaljit Singh Bhatia and Dr.Manoj Kumar

39. Simulation Study and Implementation of Single Piece Lean Line Design
    for Cycle Time Reduction - A Practical Approach
    D.Arunkumar, V.Ramesh, K.V.Sreenivasa Prasad, and T.R.Srinivas
40. Performance Analysis of AODV and DSR in MANET-A Simulative Study
   Gaurav Sethi and Manoj Kumar

41. Development of a Trusted Third Party Time-Date Stamp Server in an
   IT Security Course
   Gordon Romney

42. Simulative Analysis of operating conditions for Flat amplitude Multiwavelength
   Brillouin-Raman comb fiber laser
   Kamaljit Singh Bhatia and Manoj Kumar

43. GridARTCA: A collaborative technology environment utilizing
   Cyberinfrastructure - A preliminary report
   Kevin Franklin, Santiago Nunez-Corrales, Jose Castro, Peter Bajcsy,
   Alvaro de la Ossa, Thomas MacCalla, and Arun Datta

44. Serialized Tokens as a Means of Information Access Control
   Miles Romney IV, Todd H. Romney and Gordon W. Romney

45. There Exists a Critical Need for Role-based IDs in a University IT
   Infrastructure in Order to Meet Security Best Practices
   Gordon W. Romney

46. The Globally Distributed Ethernet Switch
   Barry Brueseke

47. Modern methods to solve Industry Problems A Lean Management
   Case Study carried out in a Processing Industry
   N. Prasanna, V. Ramesh and D. Arunkumar

48. Using Enterprise Architecture to Gain Strategic Alignment for Project
   Management in Information Technology
   Jack Elson, Jeanne Stacy, Brian Turner, and Russell F. Zimmermann
Inaugural Speaker

Dr. Dana Gibson, President of National University
(Tuesday, April 01, 8:30-10:00 am)

In July 2007, Dana Gibson became National University’s third president. Dr. Gibson joined NU from Southern Methodist University, where she had served as Vice President for Business and Finance. Dr. Gibson also served as Vice Chancellor for Administration and Finance at the University of Colorado at Denver and University of Colorado at Denver Health Sciences Center. Her experience includes service as Vice President of Finance and Administration and Chief Financial Officer of the YMCA of Denver and as Vice President for Academic and Information Services at Texas Women’s University. As a tenured Associate Professor of Accounting and Information System at TWU, Dr. Gibson also served as Speaker Pro Tem of the Faculty Senate.

Dr. Gibson earned both a Bachelor of Science degree in Business-Accounting and an MBA from TWU. She received a Ph.D. from the University of Texas at Arlington, with a major field of business accounting and minor fields of information systems and research methods. She has been a Certified Public Accountant since November 1984.

Conference Keynote Speaker

Mr. Mark S. Johnson, VP Applications Product Marketing, Oracle
"Sustaining Profitable Growth: The Phases of ERP in a Growing Company"
(Tuesday, April 1, 8:45 – 9:25am)

Mr. Johnson will discuss the challenges facing growing businesses - the (often) urgent technology needs that arise when a company experiences exponential growth. The situation demands a top tier technology platform while faced with budgetary constraints. Fast growing companies need to establish consistent, automated business processes across global operations quickly - in implementations that last for weeks not months. This requires, first, to get executive buy-in for the need to invest in technology. Whatever solution is deployed must be easy to maintain, adapt, and scale with low Total Cost of Ownership. Mr. Johnson will illustrate how Oracle technology meets the challenges of fast growing companies.
Peter Arzberger is Chair of the Pacific Rim Application and Grid Middleware Assembly (PRAGMA; www.pragma-grid.net), an open, institution-based organization of 30 institutions. PRAGMA, founded in 2002, has a mission to build sustained collaborations among researchers around the Pacific Rim by building applications on top of emerging Grid hardware and software. Connected with PRAGMA is PRIME, the Pacific Rim Undergraduate Experiences (prime.ucsd.edu) program, which provides international research and cultural internship experiences to undergraduate students. PRIME, founded in 2004, has admitted 36 students and sent students to four PRAGMA sites. Arzberger is a founding member of the Steering Committee another international activity, GLEON (http://www.gleon.org), the Global Lake Ecological Observatory Network. GLEON is a grassroots network of people, institutions, programs, and data linked by cyberinfrastructure and united by the mission to understand and predict the response of lake ecosystems to natural processes and human activities at regional, continental, and global scales. In addition, Arzberger is Director of the National Biomedical Computation Resources (http://nbcr.net), an NIH National Center for Research Resource award. NBCR’s mission is to develop computing and information technologies (e.g., end-to-end tools in cyberinfrastructure) to catalyze and facilitate biomedical research across a broad range of biological scales. He is also Chair of the National Advisory Board to the U.S. Long Term Ecological Research (LTER) network. Arzberger is the former Executive Director of the National Partnership for Advanced Computational Infrastructure (NPACI) and a former Program Officer at the National Science Foundation in Computational Biology.
Dr. Subhas C. Misra is currently a visiting scientist at State University of New York, Buffalo, USA, and is also a Director of Web World Network. He received his Ph.D. degree from Carleton University, in Ottawa, Canada, and M.S. and M.Tech. degrees respectively from the University of New Brunswick, in Fredericton, Canada, and the Indian Institute of Technology (IIT), at Kharagpur, India. Dr. Misra has several years of experience working in the academia, and the public and private sectors in research, teaching, consulting, project management, architecture, software design and product engineering roles. His current research interests include the areas of software management, software quality management, and information systems security, which are multidisciplinary, combining the fields of software engineering, information systems, organizational behavior, and technology and operations management. Dr. Misra has authored over 50 scholarly research papers and published (yet to appear) 4 books. He has won Best Research Paper Award in an international conference held in the United States. He was also the recipient of more than 15 academic awards and fellowships such as the Achievement Award at the 2007 World Congress held in Las Vegas in the United States for “contribution and dedication” to his field, and the Canadian Government’s NSERC Post Doctoral Fellowship. A mention about him and his work has also appeared in the June 8, 2007 issue of the Carleton Now newspaper. His biography has also been selected to appear in the Cambridge Blue Book, Cambridge, England, 2008. Dr. Misra is the Managing Editor of two international journals - the International Journal of Information and Coding Theory (IJICoT), U.K and the International Journal of Communication Networks and Distributed Systems (IJCNDS), U.K. He is a Guest Editor of a special issue of Computer Communications Journal (Elsevier Science). Dr. Misra is an Associate Editor of the Security and Communication Networks Journal (Wiley) and ICIC Express Letters (an international journal motivated to Innovative Computing, Information, and Control; published from Japan). He is an Editor of the International Journal of Systemics, Cybernatics, and Informatics.
A Computerworld Premier 100 Technology Leader, Norbert Kubilus has over 30 years of information systems, technology and operations experience in companies ranging from start-up to Global 1000, as well as with non-profits and government agencies. This includes having been CIO or CTO for Diamond Resorts International®, Stellcom Inc., The Leading Hotels of the World, BCM Inc., and Educational Testing Service, as well as COO of DataLEAD Communications and the Network Services Division of National Data Corporation. He was also a Technology Leadership Partner at Tatum LLC, providing technology management leadership to SMB companies primarily in San Diego, and a Faculty Fellow & Professor of Computer Science at The College of New Jersey. Five years ago, Nicholas Carr penned his Harvard Business Review article "IT Doesn't Matter" in which he argued that information technology no longer gives businesses a competitive edge. Carr called IT managers impatient, wasteful, and passive, and he criticized the hype about the so-called strategic value of IT. A year later, he expanded his thesis into a book titled Does IT Matter? Carr’s primary argument is that IT has become a commodity, like electricity, that confers no competitive advantage upon its business users. Among business leaders who have adopted Carr’s hypothesis, the result has been a de-emphasis of the role of IT in their companies. Such is the case in companies where the IT function has moved in the corporate organization to Sales or Marketing … or back to Finance and Accounting. Gartner and IDC studies tend to show that IT does matter. Companies that invest wisely in IT continue to increase revenues much faster than those that invest unwisely, too little, or not at all. A company that invests poorly in IT doesn’t increase revenues as quickly as its competitors. One that invests unwisely soon goes out of business or is acquired. This presentation addresses what technology leaders need to do to help their organizations make these wise IT investments.
Database Track Keynote Speaker

Mr. Todd Walter, CTO, Teradata R&D
"Data Warehouses"
(Wednesday, April 2, 9:00-10:00am)

Mr. Walter joined Teradata (a division of former NCR) in 1987. He designed and implemented features of Teradata, managed engineering teams and researched advanced database topics. Todd has worked directly with many customers at the leading edge of adopting data warehouse technology, and guiding the technology to solve real business problems.

Conference Closing Speaker

Dr. Shakil Akhtar
Distance Learning Techniques to Teach Networking and Databases
(Thursday, April 3, 10:30-11:30am)

While modern distance learning techniques such as video/audio conferencing, electronic communication, course online systems and course management systems work well for most classes offered in distance learning format, special techniques are used to teach IT courses that involve higher level of interactivity. The main philosophy in distance learning is the flexibility of self-paced learning that is anytime, anywhere but challenging at the same time to keep the participants actively involved in the learning process. Many students taking such classes are working professionals who seek opportunities for career advancement through professional development course and degree programs. They do not have much time to explore the technologies that support the learning system. Therefore one of the main requirements of any distance learning technique is to provide a quality education with an outstanding learner support system. The challenge becomes higher when effort is made to teach IT classes in distance learning format. The distance learning technologies commonly used nowadays are video/audio conferencing, email, and electronic conferencing. In addition, content management systems such as webct and blackboard are used as a centralized resource for class contents, students’ assessments, grades and class messages. It is possible to have some level of interactivity using these current technologies, which is very much needed for IT classes. However, a better interactivity is possible by using animated simulation, flash animation and screen capture of an activity as applied to teaching courses in programming, databases, networking, security and web design within the field of IT. The proposed presentation would cover such technologies with examples.
Abstract

Design patterns shall support the reuse of a software architecture in different application domains as well as the flexible reuse of components. In this paper, we propose design patterns for meta-search engines. We also introduce design patterns for common components of meta-search engines e.g. query interface generator, information extraction, result merger and result ranker. Presented design patterns for meta-search engines and their components are reusable, extendable and flexible. These design patterns accelerate the development process in meta-search domain and other related domains. Moreover, it promises higher quality of developed solutions. These design patterns also provide developers with a shared vocabulary for easy communication.

Keywords: meta-search, specialized search engine, design patterns

1. Introduction

Internet is flooded with information and contains hundreds of Web sites with thousands of topics in which searchers get lost while searching a topic. Second most popular activity in the Internet after e-mail is “search” [1]. One Microsoft researcher says: “Estimates are that information workers spend as much as 30 percent of their time searching for information, at a cost of $18,000 each year per employee in lost productivity” [2]. Primary search tools i.e. search engines, subject directories, social network search engines are available for the searchers but these are not sufficient to meet the requirements of users and are unable to provide the desired results. They have limited coverage, cannot locate high quality information from the invisible Web stored in accessible databases because of technical limitations or because of exclusion from the indices and the returned search results consist of long documents. Searchers have to navigate through long documents to find the relevant information from these long documents. One of Microsoft reports says: “people search an average of 11 minutes before they find what they are looking for” [3]. Subject directories are organized on the Internet sites by subject and users choose a subject of interest from the list of subjects. Subject directories depend on human editors for listings. If the description of a site is not specific enough then search may be unsuccessful. Social network services allow people with shared interests, hobbies, or causes to come together online. Social network search engines (del.icio.us, digg, reddit) are a class of search engines that use social networks to organize, prioritize, or filter search results. Social bookmark sites, allow Internet users to share content they like best with others for searching and viewing. Users can assign tags to their favourite Web pages. But the problem with such type of tag-based systems is that there is no set of controlled vocabulary for tags and no standard for tag structures. There also
exist spelling errors and multiple meanings of tags. Some users are also misusing tags to make their Web sites more visible [4].

To overcome all these problems with traditional search tools, “meta-search engine” are proposed as an choice for specific topic search. Meta-search engines (MSE) also known as multi-threaded engines, do not necessarily maintain their own listings/databases, but send the user’s query simultaneously to other search engines, Web directories or to deep Web, collect the results, remove the duplicate links, merge and rank them according to their own algorithm in a single list and display it to the user. Meta-search engines provide fast and easy access to the desired search, because they can search from multiple search engines simultaneously and save the precious time of the searcher. Meta-search engines provide a broader overview of a topic as compared to traditional search engines and increase the coverage of Web by combining the coverage of multiple search engines. Querying multiple search engines is more scalable then the centralized general purpose search engine. Meta-search engines have the ability to search the invisible Web too thus increasing the precision, recall, and quality of results. They make the user task much easier by searching and ranking the results from multiple search engines [5].

A lot of research is in progress in the field of developing configurable meta-search engines in different domains i.e. jobs, hotels, flights, news, research papers and real estate etc. It has been observed that developing a configurable meta-search engine in any domain is a tedious and time consuming task. Every time developers have to start the development process from scratch. It is desirable to have a reusable and flexible design for meta-search engines. After a detailed study of meta-search engine development research, we identified different processes and components for meta-search engines that meet specific requirements. In this paper, we propose reusable and flexible design patterns for meta-search engines and their components so that they can be reused in several times after some modifications. The design patterns of meta-search engine i.e. result ranker can also be reused in other application domains.

The rest of the paper is organized as follows. Section 2 describes the related work in the development of meta-search engines, design patterns and frameworks for domain specific search engines from a design pattern perspective. The discussion of different solutions motivate also the rationale for the design pattern. Section 3 provides an overall architecture for meta-search engines. Section 4 contains design patterns for meta-search construction and usage as well as important common components of meta-search engines. Finally, sections 5 concludes our work.

2. Related Work

In object-oriented systems there exist recurring patterns of classes and communicating objects. These patterns provide simple and elegant solutions to specific design problems and make object-oriented designs more flexible. In [6], Gof (Gang of four) describe 23 of the most common patterns in detail. These design patterns can help developers to structure their own specific applications and give them a common vocabulary to describe design concepts, rather than particular implementations [6][7].
To our knowledge, [8] is the only research that describes the framework for domain specific search engines from design patterns perspective. [8] also present design patterns for some components of domain specific search engines.

The Web Database Metasearch Engine project is developing technologies for providing integrated access to Web databases. Important phases for interface integration are automatic schema matching, schema integration and data integration. [9] use meta-information from Web search interfaces and present a two-step clustering based approach i.e. positive match based clustering and predictive match based clustering for schema matching.

MetaQuerier project (http://metaquerier.cs.uiuc.edu/) explores and integrates the query databases that are not visible to the traditional crawlers. Main components of MetaQuerier are MetaExplorer and MetaIntegrator. MetaExplorer discover sources on the deep Web and builds a search engine of Web databases. Moreover, MetaExplorer also creates models to represent discovered databases and develop wrappers to automatically extract schema details. MetaQuerier applies a holistic schema matching approach for schema matching. Holistic schema matching approach is used to identify simple 1:1 matching and complex 1:n or m:n matchings [10] [11].

In our previous research [12][13], we integrate job portals by meta-search and use a domain ontology for schema matching, schema integration, and data integration. The domain ontology is also used for information extraction from the result pages returned by various search engines. In [14] we have applied meta-search in the domain of accommodation search. Here the extraction from Web sources is based on Web service interfaces, requiring not all components that are required in the job domain. In a further project we address the search for orders in the logistics domain.

[15] introduce some techniques to automatically extract search result records (SRR) from dynamically generated HTML result pages. They present a tool ViNTs (Visual Information aNd Tag Structure) that utilizes visual content features and HTML tag structure of HTML result pages for the automatic wrapper generation of any given search engine. [16] presents techniques for supervised wrapper generation and automated Web information extraction. They implemented these techniques in a system called Lixto. Lixto provides a visual, interactive and convenient user interface for the creation of semi-automatic wrapper programs. The Lixto wrapper generator consists of modules i.e. navigator, extractor and visual developer. Lixto wrapper generator translates required piece of information from HTML pages into XML. [17] proposes an algorithm MDR (Mining Data Records in Web Pages) to mine data records in a Web page automatically. Authors claim that their data mining technique is able to mine both contiguous and non-contiguous data records.

[18] investigate result merging algorithms for meta-search engines, to merge results from different search engines into a single ranked list and state that merging based on titles and snippets of retrieved results can outperform other approaches based on full document analysis. They present five algorithms i.e. TopD, TopSRR, SRRSim, SRRRank and SRRSimMF for merging results into a single ranked list. Search result records (SRR) contain URL, title, and summary (snippet) of the extracted document. TopD algorithm uses the top document while TopSRR algorithm use top SRRs to compute the search engine score. The SRRSim algorithm
computes similarities between SRRs and query. SRRRank algorithm rank SRRs using more features like location of the occurred query term or total number of occurrences of the query term in the title and snippet etc. SRRSimMF algorithm computes similarities between SRRs and query using more features.

3. Meta-Search Architecture

This section describes the overall design of a meta-search engine. We identified that there are two processes in meta-search engines i) the meta-search engine creation process and ii) the meta-search engine usage process. Figure 1 and 2 shows the main components involved in the meta-search engine creation and usage process.

![Fig. 1: Meta-Search Creation Process Components](image)

Meta-search engine creation process work as follows. First of all, a developer specifies preferences like for which type, country or geographical area a meta-search engine is required by a preferences collector component. After getting preferences the search engine selector component will be activated and search engines meeting the preferences of the developer will be selected. Next, the Interface extractor component derives and analyse attributes from Web search interfaces. Then, an XML Schema generator component creates XML schemes for every search interface. Finally, a query interface generator component matches and integrates different XML-Schemes to have a single query interface for the meta-search engine.

When a user/seeker sends a query from the “meta-search query interface” to the meta-search engine, the components involved in meta-search engine usage process (Figure 2) will be activated and works as follows. Queries from the query interface are dispatched by a query dispatcher component and result pages with lists of results are collected. The information extractor component is responsible for extraction of records and the results from the result pages. Next all the identified results are merged together by a result merger component and stored in a
database or in XML format. Duplicate results are removed by a duplicate result eliminator component and finally results are ranked by a result ranker component.

![Fig. 2: Meta-Search Usage Process Components](image)

4. Design Patterns for Meta-Search Engines

This section contains the designed patterns for both processes and the main components of meta-search engines. Figure 3 and 4 describes our meta-search processes by using an abstract factory design pattern. “Abstract factory pattern provides an interface for creating families of related or dependent objects without specifying their concrete classes. It defines the interface that all concrete factories must implement, which consists of a set of methods for producing products” [6][7]. The meta-search abstract factory can produce any type of meta-search engine, as long as it gets proper set of directions, called factories. A factory design pattern will build a certain type of meta-search engine, depending upon the type of factory. It is clear from the abstract factory pattern in Figure 3 and 4, that families of job meta-search objects or hotel meta-search objects can be created from a developer or a user perspective. Each type of meta-search engine has the same overall structure that all meta-search engines share in common i.e. interface extractor, XML-schema generator, query interface generator, information extractor, result merger, duplicate result eliminator and result ranker etc. New factories for flights or real estates search etc. can be added easily. Using the meta-search abstract factory pattern, we do not need to worry about what kind of meta-search we are building. An abstract class may contain a default method i.e. a simple ranking algorithm, that can be used by every concrete meta-search engine, but will be refined if more specific ranking is required. Job-MSC-Factory in Figure 3 represents a Job-Meta-Search-Creation-Factory and Job-MSU-Factory in Figure 4 represents Job-Meta-Search-Usage-Factory. Few parts in Figure 3 and 4 are not drawn for the sake of saving space.
Fig. 3: Abstract Factory Pattern for Meta-Search Engines Creation Process

Fig. 4: Abstract Factory Pattern for Meta-Search Engines Usage Process
Below are some design patterns for common and important components of meta-search engines.

4.1. Query Interface-Generator

The query interface generator component of a meta-search engine is responsible for schema integration, data integration and then production of query interface for meta-search engine. Integration of interface schemes is divided into two parts i.e. schema matching and schema merging. During schema matching, semantic correspondence between interface attributes is identified and each schema is translated into a single schema for the query interface. During data integration, the values of different attributes for the user interface are determined. It is required that values are semantically unique and compatible with the local values. Different methods have been proposed for schema and data integration by using i) domain ontology [12][13], ii) clustering approach [9] or iii) holistic schema matching approach [10][11]. Different approaches for schema matching meet specific requirements according to the specific context. So developers should have a facility to choose one of the above mentioned approaches according to the specific requirement. It is required that new algorithms for schema and data integration comply with the same interface. These algorithms can easily be introduced with less effort and without changing the other code.

![Strategy Pattern for Meta-Search Query Interface Generator](image)

To meet above requirements, the strategy design pattern as shown in Figure 5 is used for the query interface generator. “Strategy design pattern defines a family of algorithms, encapsulate each one, and makes them interchangeable. Strategy lets the algorithm vary independently from clients that use it” [7]. The design pattern in Figure 5 implements that the Meta-Search-Query-Interface-Generator is a class that is responsible for schema and data integration of different search engines and Schema-And-Data-Integrator is an interface. Integrating strategies is not implemented by the class Meta-Search-Query-Interface-Generator. Instead, they are implemented separately by sub-classes of abstract Schema-And-Data-Integrator class. Sub-classes of abstract Schema-And-Data-Integrator class implement different integrate strategies i.e. Integrate-By-Ontology, Integrate-By-Clustering and Integrate-By-Holistic-Schema-Matching. To switch schema and data integrator strategies, each meta-search engine calls the integrate method that it prefers. If a developer wants to add a new schema and data integration algorithm into the system, this can be done easily by implementing a new class using the Schema-And-Data-Integrator interface.
4.2. Information Extractor

The *information extractor* component is responsible for extraction of results from the result pages. It consists of a *Record collector* component and a *Result Collector* component. The *record collector* component is responsible for identification of the record section from the result page i.e. list of jobs, table with flights etc and *Result collector* component is responsible to extract the exact fields i.e. job salary, hotel price etc, from the identified record section. Information extraction research shows that information extraction from different websites is often performed by using wrappers. Wrappers can be constructed manually, semi-automatically and automatically for record section identification. For identification of record section, different approaches i.e. i) automatic wrapper generation as ViNTs [15], ii) supervised wrapper generation as Lixto [16] or iii) data mining approaches [17] can be utilized. After identification of record section, a meta-search *result identifier* component is utilized for the extraction of exact results by using i) a domain ontology [12] [13] or ii) Lixto information extraction tool again [16].

For schema matching and merging it is required to normalize the terms like “Posted Date” into “Post Date” or “Job Types” to “Job Type”. Stemming algorithms i.e., Porter’s stemming algorithm can be utilized for term normalization process. A stemming algorithm is a method to convert word to their related form i.e. root, stem, or base. The stemming process is useful to find similar terms by only considering the word stem in search engines, natural language processing, and text processing [19].

The strategy design pattern is utilized for the information extraction component of the meta-search engine. The design pattern in Figure 6 shows that the Meta-Search-Information-Extraction class is responsible for information extraction from different result pages. Result-Collector and Record-Collector are interfaces. Sub-classes of Record-Collector abstract class implement different record identification strategies i.e. Record-Identifier-By-Automatic-Wrapper, Record-Identifier-By-Ontology, Record-Identifier-By-Lixto.
Identifier-By-Lixto and Record-Identifier-By-Data-Mining. Sub-classes of Result-Collector abstract class implements two result identification strategies i.e. Result-Identifier-By-Ontology and Result-Identifier-By-Lixto.

4.3. Result Merger

Result extracted from different search engines need to be merged and then stored for future use. Results can be stored in a database or in XML format. We re-used result merger design pattern for result merger component from [8] with small changes.

![Fig. 7: Abstract Factory Pattern for Meta-Search Result Merger](image)

The abstract factory pattern is used for the result-merger component of the meta-search engine (see Figure 7). “Meta-Search-Result-Merger” abstract factory defines the interface that all concrete factories i.e. Database-Factory and XML-Factory must implement, which consists of a set of methods for merging results. The concrete factories i.e. Database-Factory and XML-Factory implement the different product families i.e. Create, Query etc. To merge results, the client uses one of the factories and each factory knows how to create the right object for the right merging process. Few parts in Figure 7 are not drawn for the sake of saving space.

4.4. Result Ranker

A result ranker component ranks the results according to user preferences. Rank preferences can vary according to personal choices or meta-search engine type (i.e. job, hotel). For example, a seeker may want to rank the flight results according to the price, hotel according to the nearest location or job according to the query relevance. For merging results into a single ranked list according to the query relevance alone of the algorithm from TopD, TopSRR, SRRSim, SRRRank,SRRSimMF [18] can be used.
A strategy design pattern is used for the result-ranker component of the meta-search engine (see Figure 8). The Meta-Search-Result-Ranker class is responsible for ranking of results. Sub-classes of the Rank-Algorithm abstract class implement different ranking strategies i.e. Rank-By-Price, Rank-By-Location, and Rank-By-Relevance.

5. Conclusion

In this paper, we presented design patterns that can structure complex meta-search engines construction processes and provides us with flexible design strategy. We introduce design patterns for meta-search engine construction and usage as well as its components.

The difference between our and existing work is that we also introduced design patterns for i) a query interface generator component of meta-search engines that is required for user convenience and ii) information extraction component for information extraction from search engine’s result pages.

The reusable design patterns for meta-search components can be reused several times not only in meta-search domain but also in some different application domain after some modifications. Design patterns for meta-search engines are flexible enough and have facility to add or remove features for different components with minimum effort in future. These design patterns can speed up the development process for new developers as they do not need to rediscover the design problems and can save time. These design patterns for meta-search also enhance communication between project team members by providing them with shared vocabulary and provide a way to alter or extend some part of the system independently of all other parts.

The extend of reuse of software and software designs is difficult to evaluate since a reuse typically occurs years later after the design process. Moreover, the critical aspect is the reuse of software by developers not involved in the original design.
We have designed the design patterns with three applications in mind (job search, accommodation search, and search for transport orders). Moreover we have analyzed different algorithms described in the literature that can be used in meta-search.

References

[6] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995
[7] Eric Freeman, and Elisabeth Freeman, Head First Design Patterns, O’Reilly, 2004
APRIORI_PROCREATION- TOOL for Optimizing WEB PERFORMANCE

Aruna Jain
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra, Ranchi

Abstract: Mining Association Rule is one of the common forms of Data Mining, in which Frequent Item sets are retrieved efficiently. In this paper, we review the APRIORI class of Data Mining algorithm for solving the Frequent Item set counting problem and develop a new algorithm, APRIORI_PROCREATION that reduces the redundancy of Candidate Item set generated directly from Frequent Item set by narrowing the Breadth First Search Space and hence reducing its execution time leading to enhancement in mining efficiency of classical APRIORI. To validate our claim, we test our algorithm against transaction database containing Web Logs of www.ddppl.in, which shows that our approach improves the performance of Web Server considerably in terms of Cache Hit Ratio and Bandwidth Ratio by reducing the Execution Time of the web request.

Introduction:

The expansion of the World-Wide-Web (WWW) has created an unlimited opportunity for users to disseminate and gather information online. As more and more data are becoming available on the servers, there is need to make searching comprehending, and retrieval of more sophisticated and dynamic semi-structured information’s(data) stored on the Web efficiently[1]

This necessitates study and analysis of the Web-user behavior to serve the users better and increase the value of an enterprise. One important aspect of this study is the Web-Log data which embodies much of web users’ browsing behavior. From these Web Logs, we can discover patterns that predict the users’ future requests and store them in anticipation of their future requirement thereby reducing response time and improving the server performance [2]. This technique of data mining is employed to discover new information in terms of patterns or rules from vast amounts of data by building predictive and descriptive models. A predictive model is used to explicitly predict values whereas
descriptive model describes patterns in existing data, which provide information such as identifying different customer segments. Web predictive model developed by [3] shows that data mining based caching enhances hit rate and weighted hit rate significantly.

An important class of data mining problem is mining Association Rules from Web Log data. Association rule is a data mining technique which discovers strong correlation among data to understand user access patterns and improve the website design (provide efficient access between highly correlated objects, better authoring design for pages, etc.) and enable better marketing decisions (better customer classification and behavior analysis, putting advertisements in proper places, etc.)(4,5,6). Web Log analysis described in [7] is dedicated for improving Web Server Performance. The implemented algorithm is based on psychological human memory retrieval research which collects past access patterns and predicts further user action. It is observed that the Association Rule based Pre-fetching model has better predictive value than Site Structure Pre-fetching model and provides a good cache hit ratio without much addition to traffic load.

Association Rule Mining finds the set of all subsets of items/attributes that frequently occur in database records/transactions and extracts rules on how a subset of items influences the presence of another subset. Association Rule can be expressed as:

\[ A \Rightarrow B \ [S, C] \]

where A and B are sets of items; S is the support of the rules, defined as the rate of the transactions containing all items in A and all items in B i.e. Support (A \(\Rightarrow\) B) = \(P(A \cup B)\) and C is the confidence of the rule, defined as the ratio of S with the rate of transactions containing A i.e. \(P(B/A)\).

Support and confidence are measures of the interestingness of the rule. A high level of support indicates that the rule is frequent enough for a business to be interested in it. A high level of confidence shows that the rule is true often to justify a decision based on it. Minimum support / confidence required for a rule to be reported is its threshold value.

One of the main attributes needed in an Association Rule-mining algorithm is Scalability, the ability to handle massive data stores. As a result fast and efficient Association rules are required to handle increasing number of transactions in real world databases. These rules are able to discover related items occurring together in the same
transaction. Since these transaction databases contain extremely large amount of data, current Association Rule discovery techniques try to prune the search space according to the support for the items under consideration.

In the application domain, items correspond to web resources, while transactions correspond to user sessions. Thereby, a rule such as res₁ ⇒ res₂; mean that if res₁ appears in a user session, res₂ is expected to appear in the same session, though possibly in reverse order and not consecutively. The basic algorithm called Apriori Algorithm for finding the association rules was proposed in 1993 [8] and later modified by [9,10] uses Breadth First Search, Bottom Up Approach and performs well when the Frequent Items are short and thus it is easy to implement when the support required is high as it leads to a smaller number of frequent items. But for larger number of frequent items it generates huge set of candidate items leading to high memory requirement and more searching time. Our work therefore focuses on developing an efficient algorithm of mining frequent item sets for association rules with Procreation count .We modify Apriori by using Procreation Count of Frequent Item sets at a level, which is related to Support Count of Candidate Item sets at a next level. Our modification leads to reduction of total number of Candidate Item sets by reducing number of rows in a transaction database that reduces the Cardinality of candidate item sets to improve efficiency of finding frequent item sets. The organization of the paper is as follows –

In Section 2, we give analysis of APRIORI Algorithm and define terms, which are needed to understand our algorithm. In Section 3, we describe our new algorithm, APRIORI_PROCREATION. Section 4 gives the result of implementation of this algorithm on the data set of www.ddppl.in and finally Conclusions are drawn in Section 5 with a comment on future research opportunities in this area.

2 Analysis of APRIORI Algorithm:

A set of items containing k-items is referred as a k-item set. Support count or frequency support count or frequency of an item set is the number of transactions that contain the item set. An item set satisfies minimum support if the occurrence frequency of the item set is greater than or equal to the product of minimum support and the total number of transactions in D. The number of transactions required for the item set to
satisfy minimum support is referred to as minimum support count. If an item set satisfies minimum support, then it is a frequent item set.

Discovering Association Rules is divided into two steps. In the first step, all sets of items with \( \text{Support} \geq \text{Minsup} \) called the Frequent Item sets are found. In the second step the Frequent Item sets are used to discover the Association Rules. Mining Frequent Item set is a fundamental and essential operation in Data Mining application including discovery of Association Rule and Sequential Rules and is difficult to resolve. We try to solve and optimize this search for Frequent Item set in the proposed algorithm.

2.1 Finding Frequent Item sets with procreation Count:

In Apriori algorithm, we generate candidate item sets \( C_{k+1} \) from frequent item sets \( L_k \) in the \( k^{th} \) mining loop. A large number of redundant candidate item sets are generated in the process because the support of some \( k + 1 \) item sets which are supersets of frequent item sets in \( L_k \) is smaller than Minimum Threshold value. We want to make cardinal number of candidate item sets \( C_{k+1} \) near to cardinal number of \( L_{k+1} \). So, when we count the support of candidates in \( C_k \), we also calculate procreation of those candidates and if procreation of a candidate in \( C_{k+1} \) is less than minimum threshold then we do not include that in \( L_{k+1} \). Thus we are reducing cardinal number of item sets in each level and ultimately reducing the search space to a significant extent in the final level. Hence, volume of the database will be reduced and mining efficiency will be improved.

2.2 Procreation count of frequent item sets:

We define procreation count of frequent item sets and then use this procreation count along with already existing support count to modify Apriori algorithm to increase efficiency of Apriori by reducing Breadth First Search space.

Let \( 1 = \{i_1, i_2, \ldots, i_n\} \) be a set of items. Let \( D \), the transaction database, be a set of transactions, where each transaction \( t \) is a set of items. Thus, \( t \) is a subset of \( 1 \). Let \( x \) be a frequent item set and \( x \cup \{i\} \) is next level procreated element, \( x \cup \{i\} \) is generated through candidate generation procedure used in Apriori algorithm. If \( D \) contains total \( n \) transactions and if total \( m \) transactions contain \( x \cup \{i\} \) then

\[
\left( \frac{m}{n} \right) \times 100 = P
\]
P is called Procreation of $x$.

2.3 Relationship between Procreation and support of an item

If procreation count of frequent item set $x$ in ordinal item sets tree is less than minimum threshold value, then support count of all children $x$ denoted by $xU \{i\}$ is also less than minimum threshold value.

Vegetal item sets:

If procreation count of frequent item sets $x$ is greater than or equal to minimum threshold value, then $x$ is called vegetal item sets.

3 Algorithm APRIORI_PROCREATION for finding frequent item sets:

Let $U_k$ and $U_2$ are two infrequent item sets, which are used to decrease volume of database. Counter 1 is used to count the procreation in every loop. Algorithm generates frequent 1 item sets, frequent 2 items sets and finally frequent k item sets.

Transactional database and Minimum Threshold values are the two inputs of the algorithm. With a given minimum confidence threshold and a minimum support threshold, the problem of mining association rules is to find all the association rules whose confidence and support are larger than the respective thresholds. The higher the threshold value, the more precise the prediction and the less aggressive prefetching. In our earlier work [7], we have defined 25% minimum support and 60% minimum confidence as good breaking point.

Begin ;

1: $F_1^k = \{\}$   $F_1^k$ is a set of frequent item sets.
   $F_2^k = \{\}$   $F_2^k$ is a set of vegetal item sets.
   $U_k = \{\}$   $U_k$ is a set of k-infrequent item sets.
   Result = \{\} Result is a set of final outputs.

Initially there are empty sets */

2: Initialize $k = 1$;   /* we start from first level */

3: $A_k = \{\text{frequent 1 – itemsets}\}$   /*$A_k$ is a set of item sets which contains all 1- item sets */

4: If ($A_k \neq NULL$)   /* if $A_k$ is empty for some $k$ then algorithm ends*/
   repeat Step 5 to 17

5: Initialize $j = 1$ ;   /* We will consider each transaction $T_j$ one by one
repeat Steps 7 to 12 until j > m

m is total number of transactions in our database*/

6: If
   cardinality of T_j < k
   Then
   Remove T_j from database

7: If (1k belongs to Uk ^ cardinality of T_j = k+2 ^ T_j-1k belongs to Uk-(k-2) /*
   Then
   k-itemsets T_j
   remove T_j and all it’s subsets

8: If
   T_j belongs to Uk-1
   Then
   Remove T_j from database

9: for (all i, where i belongs to T_j)
   support_counter++; /*
   i is an item in transaction

10: for (all i, where i belongs to T_j ^ cardinality of T_j > k)
    procreation_counter++;;

11: j++

12: for all I do
   if
      support_counter of I >= minimum threshold value
   Then
      F_1k = F_1k U {I}

13: for all I do
   If
      procreation_counter >= minimum threshold value
   Then
      F_2k = F_2k U {I}

14: Uk = Ak - F_1k;

15: Result = U F_1k

16: k++

17: Ak = gen_candidate_itemsets with the F_2(k-1)

End:
In step 6 of the algorithm, we consider transactions which have minimum cardinality during first iteration. In the next iteration we consider transactions having minimum cardinality 2 and so on. When we are in kth loop, all item sets of cardinality k-1 are already included in the set of frequent F-1k or to the set of infrequent itemsets Uk. So now we can remove Tj.

In step 7 we have an itemset which is known to be infrequent and whose length is 2 less than the length of transaction Tj in database. If after removing the infrequent itemset from the transaction, remaining set of items is still infrequent then remove Tj and all its possible subsets.

i is an item in transaction. If we find a transaction Tj which has already been included in infrequent itemsets then delete Tj because we cannot procreate any new frequent itemset from Tj.

In step 9 we increment support count of each itemset belonging to Tj because the support for an itemset is probability of inclusion of the itemset in a transition Tj.

In step 10 if the cardinality of item set is greater than present k value, the procreation counter is incremented by 1. This procreation counter is used to determine whether itemsets are vegetal or not.

Itemset having support count greater than minimum threshold values are included to two set of frequent item sets F1k in step 12. Similarly itemsets having procreation count greater than minimum threshold value are included to set of vegetal itemsets in step 13.

We remove all frequent itemsets from candidate itemset to get a set of infrequent itemsets in step 14 and take union of all frequent itemsets generated so far to get final result in step 15. Before going to next iteration, k is incremented.

Since candidate itemset Ak for current iteration should include all frequent and vegetal itemsets of previous iteration, so we modify Ak in step 17.

4 Implementation Methodology And Results: The proposed implementation methodology approach is primarily based on knowledge discovery of server access patterns through data mining. The data mart is populated starting from raw web/proxy server log files of www.ddppl.in, a leading travel industry publishing house which has offices in Delhi, Mumbai, Sri Lanka and Dubai. This site is not big in terms of contents
but then it has several directories and html files. The data mart population consists of a number of preprocessing and coding steps that perform data selection, cleaning and transformation.

The data mart has been implemented as a relational database, using Microsoft SQL Server 2000 Beta2. The process of data preparation and data mart population have been designed using SQL Server 2000 Data Transformation Services (DTS), a tool that allows to specify import / export / transformations processes of data through text files, databases or applications. Important tasks include field extraction, URL normalization and hash coding.

The quality of a web caching scheme is evaluated on the basis of Cache Hit Ratio, Bandwidth Ratio and Latency depending on the resource that is being focused on. We measured the Cache Hit Ratio which is defined as the ratio of number of objects served from cache to the total number of objects requested. A high hit rate reduces server load by reducing the number of requests a server must process. We also measure Bandwidth Ratio, which is defined as the ratio of total number of files fetched from the server to the total number of requests.

We extracted session IDs and the list of URLs requested in each session from the server log and assigned a unique number to each URL.

4.1 Results:

Some of the key findings are:

1. To justify our statement that APRIORI_PROCREATION reduces the searching space for Breadth First Search by lowering the scale of candidate items as compared with classical APRIORI, we plotted a graph of Execution Time Vs. Minimum Support and observed that the new algorithm is more effective in decreasing data size leading to reduction in execution Time as shown in Figure1. This fall in execution time is
prominent when the minimum support level changes from 20 to 30 in which case Apriori Procreation decrease

2. To study what effect this reduction in Execution Time makes on Quality of a Web Caching Scheme, which is evaluated in terms of Cache Hit Ratio and Bandwidth Ratio, we plotted these parameters against Minimum Support. Cache Hit Ratio is the ratio of number of pages found in Cache to the total number of pages requested. From Figure 2, which shows a plot of Cache Hit Ratio against Minimum Support we found
that Cache Hit Ratio is increased from 0.35 to 0.50 at a support level of 10. However, for higher range of Support (between 20 to 30), Cache Hit Ratio increases at a lesser rate implying that aggressive prefetching beyond a certain point does not yield proportionate benefit.

Figure 3 shows a plot of Bandwidth Ratio against Minimum Support. Bandwidth Ratio is defined as the total number of files fetched from the server divided by the total number of requests. We note that as far as Bandwidth is concerned though the number of files prefetched increases with decrease in Support the Bandwidth Ratio decreases marginally because low support generates large number of rules thus causing more files to be prefetched.

These figures would vary for different weblog datas from different sites because it depends on how strongly various pages are related to each other and how good the association rules are that we find for those sites.

5 Conclusion: We have studied how the knowledge discovered through Web Log Mining can be used to optimize Web Server Performance thus improving user perceived latency on the World Wide Web. Implementation of our new algorithm,
APRIORI_PROCREATION on the Web Logs of www.ddppl.in shows that execution time for serving the user request is reduced as compared to classical APRIORI. This reduction in execution time brings proportionate increase in quality of Web Caching Scheme evaluated in terms of performance measure such as cache Hit Ratio and Bandwidth Ratio.

From our experience in this study, we believe that there are many open research issues. One direction would be to obtain a predictive model based on Sequence Rules and compare its performance with that of model based on association rule with respect to both Classical APRIORI and APRIORI_PROCREATION.

REFERENCES:


5. J. Han, J. Pei, “Mining Frequent Patterns by Pattern-Growth Methodology and Implications”, ACM SIGKDD Explorations (Special issue on Scalable Data Mining Algorithms), December 2000.

6. J. Fong, J. G. Hughes and J. Zhu, 2000, Online web mining transactions association rules using frame metadata model..


Formal Privacy Ontology, a Definition
Position Paper

Waël A. Hassan

University of Ottawa, Ottawa, Ontario, 800 King Edward, K1N 6N5
whassan@site.uottawa.ca

Abstract. Legislative compliance in the area of information management is of growing concern to institutions and enterprises. The Privacy legislation often adds interdependence and possibly ambiguities that need to be resolved prior to enforcement, audit, or implementation. Privacy must communicate requirements in a precise language understood by security information technologists and architects. Organizations must address privacy requirements to meet (a) their legal obligations, (b) their contractual obligations and (c) the expectations of the data subjects and client groups. A privacy ontology provides a basis for collecting engineering requirements specifying privacy governance rules, and in this sense, the represents the privacy meta rules for an organization.

Keywords: Privacy, Formal, Ontology, Policy, Governance.

1 Introduction

Computing systems are taking an increasing role in validating privacy compliance [4] [8] [14]. Validating legal compliance to privacy aims at ensuring that an enterprise has the necessary processes, roles, technical measures, and rules to govern and protect personal information. Canadian privacy laws stem from the Canadian Standards Association’s “Privacy Code” otherwise known as the ten principles. These principles are accountability, identifying purposes, consent, limiting collection, limiting use, disclosure, and retention, accuracy, safeguards, openness, individual access, and challenging compliance. These principles govern collection, use, retention, and disclosure of personal information. Canada being a federal state has several layers and areas of jurisdiction (federal, provincial / territorial, and municipal). Organizations operating in Ontario, for example, may be subject to the Personal Information Protection of Electronic Documents Act (PIPEDA)[19], the Privacy Act, Freedom of Information and Protection of Privacy Act (FIPPA) [22], the Municipal Freedom of Information and Protection of Privacy Act (MFIPPA) [23], and / or the Personal Health Information Protection Act (PHIPA) [24]. In this paper we propose the use of legal ontologies for the purposes of capturing consistent privacy requirements.
The privacy domain faces the challenge of requirements capturing, refinement of specifications, in addition to interaction-detection. Privacy requirements are presented at different levels of specification and using various representation styles. A legal requirement, for example, may be represented as a declarative rule, whereas enterprise policies are represented in a functional ‘if-then-else’ style. Combining rules of operational style with declarative style is a challenge. Translating requirements to specifications and implementation is a generic problem that is the major concern of Computer Science (‘CS’) research. Jackson and Zave [11] [16] argue that “requirements exist only in the environment” and domains should be described explicitly[5] independent of the system to be built. Privacy requirements introduce two further complexities; the first is enforcement, the second is context dependency. Enforcement is a challenge since people’s trust and use of data is a predicate in the underlying system. Context capturing, on the other hand, is a challenge since the applicable set of policies changes based on context of execution. These issues and others are captured in a roadmap presented by [4]. It is certain that the trend of requirement translation presented in [7] and further studied in [2] still holds:

<table>
<thead>
<tr>
<th>Table 1 Layers in Software Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Requirements Engineering</td>
</tr>
<tr>
<td>• Requirements Specification</td>
</tr>
<tr>
<td>• System specification and design</td>
</tr>
<tr>
<td>• Programming Languages</td>
</tr>
<tr>
<td>• Assembly Code</td>
</tr>
<tr>
<td>• Machine Code</td>
</tr>
</tbody>
</table>

Requirements engineering [9] has traditionally been concerned with investigating the goals, functions, and constraints of (software) systems. The process of capturing requirements can be broken down into four tasks: elicitation of information related to the problem domain; modeling of the problem; analysis of costs, completeness, and consistency; and validation with the customer [1]. These tasks pave the way to the generation of complete, consistent, and unambiguous specifications of system behaviour that are well suited for design and implementation activities [6].

There are several examples [4] of requirement engineering resulting in privacy breaches, violations, and conflicts. Therefore we classify the issues facing privacy requirements specification, translation, and conflict detection into Privacy generic problems and domain specific challenge.

Privacy domain [4] specific issues include the:
- Need for an abstract representation of privacy domain;
- Lack of precise specification models in standards;
- Need for a technique that assists in translating privacy semantics to operational constructs; and
- Need for a method that defines the integration of privacy policies into a multi-model security environment.
Generic specification issues include the inability to combine operational and declarative specifications, and the need for policy interaction detection. We suggest that ontologies can assist in the abstraction, increasing of preciseness and provide meta rules for creation of system rules from requirements or specifications.

Ontologies consist of sets of statements that describe definitions or characteristics that must be satisfied by (the ontology designer’s idea of) the “reasonable” state of the world. Formally, such statements correspond to defining terms used in logical sentences, and ontology corresponds to a logical theory. Privacy ontology will be able to offer the core abstraction to deliver such a promise. We propose the use of a decoupled ontology capturing:

- Legal Entity Ontology
- Data Entity Ontology

The paper proceeds to describe related work in Section 2. Section 3 contains the related work laying out existing ontologies and frameworks.

2. Related Work

Creation of a privacy problem as specified in the Privacy Incorporate Software Agent (‘PISA’) [15] requirements documents is rooted in the abstraction and implementation of privacy legal requirements approached in three distinct methods:

1. Existing privacy ontologies
2. Existing privacy frameworks.

2.1 Existing Privacy Ontologies

There are a limited number of privacy ontology examples, including DAML and PISA.

**DAML Privacy Ontology**
DAML developed a generic, simple and easy-to-use ontology for expressing privacy policies as well as a protocol to support matching of privacy policies. DAML distinguishes between three types of rules: Authorization, Capability, and Obligation.

**PISA Ontology**
The PISA ontology includes some meta ontology concepts such as data protection authority, privacy policy, privacy preference, and privacy principle. In addition it includes a classification of personal data into three data types. Finally, it breaks down the privacy principle into the concepts of transparency, finality, data subject rights, legal processing, and transfer.
2.2 Existing Privacy Frameworks

The current International Standards Organization (‘ISO’) and Organization for the Advancement of Structured Information Standards (‘OASIS’) documents lack a formal meta-model that captures privacy requirements. For example, Extensive Access Control Markup Language (‘XACML’), a product of OASIS framework has a language model is used to represent security policies. These standards provide recommendations on presentation and management of policies through languages such as XACML. We suggest and others agree, XACML is missing the constructs needed to capture requirements and operational semantics required to implement privacy. XACML policies, for example, relate targets (subject, action, resource) to an obligation, which is a generic solution to a security problem. Context (purpose), which would speak to privacy compliance, is not included in this specification.

Canadian Standards Committee on Privacy (‘CAC-P’) handles recommendations to the ISO. Their work is limited to business recommendations and framework without any privacy model. One reason behind the lack of initiative to create a specification model is that standards bodies view privacy as front-end to security. The private sector industry, on the other hand, has followed a pragmatic approach by adopting managed-policy governance systems. These systems combine process and technology to implement varying degrees of privacy. The military’s security work, for example, since the 1970s has focused on access models and algorithms related to program analysis.

3. Privacy Ontology

In this section we explain the goals, provide a formal definition for legal and data ontologies, in addition to suggesting potential applications of an ontology. Advanced normative systems use extensive sets of definitions to structure the domain on which they act. Companies have organizational structure that is taken into consideration in company policies. For example, employees can be characterized by roles. These definitions form ontologies, which are hierarchical data structures containing attributes for the entities in a certain domain, together with their relationships. Some literature refers to ontologies with the name of world knowledge, which they contrast with normative knowledge. The study of ontologies for legal systems is a research area in which Jurisprudence seem to be taking the greatest inspiration from Information Technology.

3.1 Goals

We believe that future work in the area of defining legal ontologies can help the following three areas: Requirements Engineering, Requirements specification, and system specification and design.
Requirements Engineering:
- Establish a common vocabulary: the ability to implement precise semantics is highly dependent on a consistent and precise vocabulary. A privacy ontology is able to provide the redefinitions of key terms, in addition to connectivity between equivalent concepts.
- Establish structure and legal dependency: In a multi-level, dual jurisdiction system, such as the case of Canadian provinces, the ability to understand legal applicability is important to applying the appropriate provision.

Requirements Specification
- Assist in identifying entities: Given that laws are entity based, a precise definition of entities removes ambiguities, it also assists in finding applicable laws.

System specification and design
- Defines policy format: A legal and data ontology should specify enterprise and institutional Meta Policy requirements, and hence they should affect policy format.
- Apply rules of priority and precedence: A legal hierarchy defines precedence and applicability. The ability for ontologies to define legal precedence is also an interesting area of research.

3.2 Semi-Formal Definition of a Privacy Ontology

In this subsection we present a formal definition of atomic and composite elements represented in the UML model. The we provide a predicate logic definition. The first definition formally presents a Legal Ontology. The second refers to a data ontology. We realize the importance of the data and policy components of the ontology. Both are instrumental in completing requirements for privacy policy definition and abstraction. As we mentioned in the motivation section, this will be future work.

Legal Representation

Fig. 1. Legal Ontology MetaModel

Figure 1 shows the Meta Model of the legal ontology using a UML class diagram. The ['Legal Entity'] class represents a super class. It is composed as UML generalization. The idea here is that Legal Entities are generalization of one another. ['Law'] Meta class is attached to some Legal Entities.
Data Representation

Fig. 2. Data Ontology

Figure 2 shows the Meta Model of a Data ontology using a UML class diagram. The ['Data'] class represents a composition relationship to other data classes. The idea here is that Data sets are compositions of one another. ['Sensitivity'] classes has a one-one relationship with datasets. The sensitivity class suggests a sensitivity level for various data compositions.

3.3 Formal Representation

We define formally a decoupled legal and data ontology. A Legal Ontology as defined in the UML MetaModel defines relationships between laws and applicable entities, whereas a data ontology separates between various classifications of data sensitivity.

Definition 1. Legal Ontology

A privacy ontology denoted by PO=(λ,ε,α). λ is a finite set of laws, ε is a set of legal entities, and α is an ordered relationship between members of λ and ε.

\[
P = \begin{cases} 
(\lambda) & \text{is a set of Laws} \\
(\varepsilon) & \text{is a set of Legal Entities} \\
(\alpha) & \text{is a set of Legal Applicability} 
\end{cases} \\
\alpha \subseteq \left\{\{u,v\} \mid \forall u \in \lambda, \forall v \in \varepsilon\right\}
\]

The super set of entities denoted by E=(ε,τ). Where ε is a finite set of entities. And τ is a set of composition relationship. For example, \{(u,v)\} ∈ τ => u is a type of entity v.

\[
E = \begin{cases} 
(\varepsilon) & \text{is a set of Legal Entities} \\
(\tau) & \text{is a type of relationship} 
\end{cases} \\
\tau \subseteq \left\{\{u,v\} \mid \forall u, v \in \varepsilon\right\}
\]
Definition 2. Data Ontology
A data ontology denoted by \( DO=(\delta, \sigma, \theta) \). \( \rho \) is a finite set of processes and \( \sigma \) is a set of ordered pairs of members of \( \rho \). We call members of \( \rho \) and \( \sigma \) processes and contained processes, respectively. The process set \( \rho \) of a process ontology \( P \) is denoted as \( \rho(P) \). Similarly, the containment relationship is denoted as \( \sigma(P) \). Finally, the set of purposes is denoted by \( \theta(P) \). A containment relationship

\[
DO = \begin{cases} \delta, \text{is a set of data elements} \\ \sigma, \text{is a sensitivity Level} \end{cases} \quad \sigma \subseteq \{ \{ u, v \} | u, v \in \delta \}.
\]

The super set of data elements denoted by \( \Delta=(\delta, \kappa) \). Where \( \delta \) is a finite set of Data Elements. And \( \kappa \) is a set of composition relationships. For example, \( \{ (u, v) \} \in \kappa \Rightarrow v \) is included in the composition of \( u \).

\[
\Delta = \begin{cases} \delta, \text{is a set of data elements} \\ \kappa, \text{is a composition Property} \end{cases} \quad \kappa \subseteq \{ \{ u, v \} | u, v \in \delta \}.
\]

3.4 Example Application

Figure 3, illustrates a legal applicability map for entities in Ontario, including individuals, corporations, government, non-profits, non-governmental organizations, and registered charities.¹

---

¹ [http://ontarioprivacyontology.googlepages.com/home](http://ontarioprivacyontology.googlepages.com/home)
3.5 Using a Privacy Ontology

We adopt the methods of constructing a privacy policies being proposed by Hassan and Logrippo to appear. At the abstraction stage, their method suggests the following process.
1. collect the list of purposes for which the entity consumes data while identifying the data without a specific purpose;
2. map the purposes to existing entity processes, note that the mapping from purpose to process is a many-many relationship;
3. identify applicable laws and specify system policies enforcing legal requirements;
4. define contractual obligations;
5. define policies using privacy constructs of (consent, collection, distribution, retention, secondary use).

Privacy design principles, expressed by the meta-rules in the proposed ontology, provide the cornerstone for organizations operating in Ontario to be privacy compliant. Yet, the creation of a support organizational culture of privacy goes beyond the requirements. Creation of policy and management processes are of great importance to the fields of information management and privacy. Other areas for research include issues relating to the responsibility for educating individuals on the exercise of their privacy rights under the legislation, analysis on the extension of privacy impact assessment requirements beyond the realm of health care, and specific procedures questions regarding the use of data sharing agreements between multiple levels of government in a pan-Canadian jurisdictional environment for the purpose of protecting informational privacy. These are questions of the broad sort, questions of education, expertise, assessments, scope and cooperation.

4 Conclusion

We are proposing a generic definition of an ontology that can be used at the requirements engineering level in the software engineering process. The definition provides a precise specification both semi-formally and formally. The specification’s applicability needs to be validated. Moreover, we intend to develop a formal analysis tool to validate its features. We believe that legal and data ontologies can set the basic definitions, describe dependencies, meta rules and rules of precedence for privacy policy construction in federal jurisdiction. It is our intent to prove and validate that the ontology is implementation independent, decouples rules from structure, in addition it is designed for change.
Acknowledgment. I am indebted to my supervisor Luigi Logrippo at the UQO Université du Québec en Outaouais for introducing me to this area of research and for sharing with me his ideas on Ontologies and formal methods in general. As well, I am thankful to Tracy Kosa for her numerous detailed comments on earlier work related to this paper, and for introducing me to Privacy Impact Assessment Models, laws in Ontario and providing me with applicable examples.

References


Formal Privacy Ontology, a Definition


A Method for Optimizing the Size of Wireless Mobile Messages

P P Abdul Haleem and M P Sebastian

Abstract—Optimization of the message size is a matter of concern in the resource limited, mobile wireless devices. As on today, no uniform standard is available for optimized wireless mobile messaging. The heterogeneous nature of mobile wireless devices makes the formulation of a uniform messaging standard a challenging task. The eXtensible Markup Language (XML), a universally accepted method for representing data, is a potential choice for this purpose. However, limitations like verbosity and the need for strict structuring make the XML format less attractive for wireless mobile applications. This paper proposes SA-YAML, a schema attached messaging format based on YAML Aint Markup Language (YAML), as an alternative for optimizing the size of mobile messages. SA-YAML makes use of the inherent properties of YAML in squeezing the message size, in addition to two phases of optimization. The evaluation results indicate the usefulness of SA-YAML over XML as a promising message format in wireless mobile devices.

Index Terms—Mobile Computing, Wireless Mobile Environment, XML, Knowledge Representation, Squeezing of Mobile Messages, Streaming of messages.

1 INTRODUCTION

Wireless mobile devices have many limitations such as power, memory, etc. These limitations in turn prevent the use of processors with high computing power in mobile devices. Limited bandwidth, increased latency and increased retransmission due to packet losses are still concerns in the wireless networking world. Formulating an efficient method for the transfer of messages in wireless mobile devices is a need of the hour. It is desirable to have a cross platform technology that can be used to achieve seamless access to the various ranges of devices and systems. The search for such an acceptable cross platform standard normally converges to XML. In Simple Object Access Protocol (SOAP), XML is already in use as a message syntax format [2]. The main factor in favor of XML is its continued use as the de facto standard to represent data over net connected heterogeneous systems. However, XML is crippled with many limitations [1] such as high verbosity (results in larger message size and vulnerability to retransmissions), strict structuring (verification of this structure makes XML parsing a heavyweight process), and high textual content (requires string parsing for further processing). Many web services make heavy use of messages for exchanging data.

This paper is aimed at proposing a standard message format, particularly for the wireless mobile environment. The standard is expected to retain the merits of the XML format, with improvements on its limitations with respect to the wireless mobile environment. This paper is organized as follows: Section 2 presents the related work. Section 3 discusses the design principles. Section 4 presents the system architecture. Section 5 gives the performance evaluation and results. Section 6 concludes the paper.

2 RELATED WORK

Several solutions are proposed in the literature to overcome the limitations of XML in wireless mobile environment. Some of these approaches are successful in tackling the limitations for narrow application. A closer look at these methods reveals certain problems when applied to the wireless mobile environment. For instance, standard compression techniques such as Millau [6], Gzip [7] and XMill [7] give good performance only with larger messages with a high redundancy rate. But it may not yield good results with short messages with less redundancy, typically used in the wireless
mobile networks. The compression/decompression layer is considered as an additional burden for the resource critical mobile devices [3]. Several alternative serialization methods ([10], [11], [12] and [13] for instance) are also proposed. As these formats remain neutral to the natural language, problems like readability issues and loss of description could be possible when these techniques are employed.

Our proposed work retains the agility of the legacy XML technology - at the same time, some of its inherent deficiencies specific to wireless mobile environment are addressed.

3 The Proposed Method

We have chosen YAML as the vehicle to achieve our goal. YAML is a light weight and computationally powerful data serialization language. The primary reason for considering YAML for our work is its low verbosity, at the same time being expressive and extensible. It is also easily readable and has a consistent information model. YAML supports stream-based processing too. YAML has a human-readable data serialization format that takes inputs from the languages such as XML, C, Python, Perl, and also the format of electronic mail as specified by RFC 2822. It is optimized for data serialization, configuration settings, logs files, Internet messaging and filtering [4].

Apart from just suggesting YAML as an alternative, we have proposed important value additions to ensure better throughput. The proposed format, christened as SA-YAML (Schema Attached YAML), is obtained in five stages. In the first stage, the message is encoded in the natural YAML format. This natural YAML format is optimized in the remaining stages. The second stage bifurcates the structure and content to create a schema definition of the message content, christened as YASchema, and the third stage verifies and validates the definitions made in the second stage. The fourth and fifth stages squeeze the message as obtained from stage II by making use of the properties of YAML and the YASchema, respectively.

4 Design

4.1 Stage I: Message encoding in YAML format

YAML streams are encoded using the set of printable Unicode characters, either in UTF-8 or UTF-16. Creation of messages is easy and can be done with any text processor. A detailed account of YAML specifications is given in [8]. The output of this stage is the message encoded in YAML natural format.

4.2 Stage II: YASchema Creation

The YAML data serialization includes both data description and content as an interwoven combination. Removing the clutter from the message can be achieved by bifurcating the structure and content. There exists a schema validator for YAML files called kwalify [5], developed by Kuwata-lab. YASchema attempts to improve upon the Kwalify schema in the following ways: (i) the YASchema is constructed as per the flow-style method of creating YAML messages. This helps to squeeze down the overall verbosity of the message, (ii) without affecting the readability, elements of YASchema and their properties are stored in the schema definition file in such a way that they require only less number of bytes, (iii) facility to define ID codes for elements of YAML are given - this arrangement significantly reduces the size of the original message, and (iv) the process of reusing the same schema across multiple sessions is envisaged - special directive is included in the YAML file to specify the schema definition to be referred.

There are three kinds of nodes in YAML - scalar, sequence, and mapping [8]. Apart from this, there can be mappings of sequences and sequences of mappings also. We evaluate the message structure to decide about the possibility of folding node contents. This is necessary as a blind conversion of the message to flow-style may not produce a convincing result. Nodes of the categories belong to sequences, mappings, sequences of mappings and mappings of sequences are marked as squeezable. In addition to this, we initiate the schema creation in this stage. Evaluation of the messages structure is used to decide the type of the schema to be constructed. Type can be of any of the three kinds - scalar, sequence and mapping. Primitive details like name of the field, its data type and a ID value are added to the schema for every member in a node. The output of this stage is YASchema together with message formatted in YAML natural format.

4.3 Stage III: Validation and Verification

This stage is used for the validation and verification of the schema constructed in stage II. This stage is used to correct anomalies, and to add more details like whether mandatory or not, expected pattern of data, default values etc. The output of this stage
is validated YASchema together with message formatted in YAML natural format.

4.4 Stage IV: Phase I Optimization

Normally, information representation in YAML is identical to that of XML. This enhances the readability of the content, but results in more number of bytes for preparing the messages. Interestingly, the so called Flow style method specified in YAML helps one to squeeze the size considerably. Reorganization of the contents in flow style yields better performance. For e.g., a complex message consisting of customer invoices having many product details each can be presented to this stage, for better conservation of message size. Stage IV scans the message contents with the help of YASchema and reorganizes the information that are identified as fold-able (in stage I) to flow style. The output of this stage is semi optimized message formatted in YAML together with YASchema.

4.5 Stage V: Phase II Optimization

This stage concentrates on further optimizing the messages, making use of the enhancement to YAML, namely YASchema. All primitive details regarding every element in the message are included in the schema at the schema creation stage itself. Special ID codes are provided for each element. Equipped with these information, a scan over the message encodes the elements with the ID codes. The ID codes can be obtained from the accompanying schema (Figure 1). If a sender likes to reuse a schema that was sent earlier, the schema to be referred can be specified in the message itself. This makes the decoding easier. Such re-usability of schema over multiple streams is not possible for YAML in its natural format. The output of this stage is the optimized message formatted in YAML together with YASchema.

5 Testing and Performance Evaluation

The testing is done for five categories of messages including short (simple messaging format with only text), small (a single record consisting of String, Float and DateTime types of data), medium (details of 25 customer records), large and composite (these two categories can be invoices with many records). These messages are tested for five types of formatting which include YAML natural format, semi optimized format (after applying Phase I of squeezing), optimized format (after applying Phase II of squeezing), XML, and SOAP. XML and SOAP formats are used as benchmarks. Three measures that are indicative of the resource usage and viability of the proposed message format are identified for performance evaluation. These measures are Schema Size, Message Size and Transmission Speed.

5.1 Schema Size

For each category of messages the schema definitions are created as per the specifications of Kwalify and YASchema. The gain in number of bytes is tabulated in Table 1. YASchema offers considerable optimization for all types of messages. The Short category yields the maximum gain. The Large and Composite categories (with gains 53.23% and 54.78%, respectively) are profitable for repeated occurrences of records. In the Medium category the gain is over 50%. This is also impressive taking into consideration of the re-usability of the schema over multiple streams.

5.2 Message Size

The message sizes obtained in each of these categories are listed in Table 2. The sizes of the YASchema (developed in stage II) in each category are also included. The performance advantage in squeezing down the message size can be derived from the table. For example, the gain in message size (in bytes) using YAML natural format, over messages formed in XML for Short, Small, Medium, Large, and Composite categories are 46, 153, 2816, 5024, and 19277 bytes, respectively.

The results for semi optimized message format are as expected. The gain in message size for optimized format is tabulated in Table 3. The performance gain for the five categories of messages compared to XML are 74, 250, 7217, 12951 and 44927 bytes, respectively. This measurement excludes the size needed for schema (as shown in Figure 2). For small message types, especially for Short and Simple, squeezing beyond YAML natural format is not profitable (if these messages need to send only once), as seen from Figure 3. When the overhead needed to maintain the schema for XML is also taken into account, the optimized format outperforms XML in all categories. Even though there is an apparent marginal gain (e.g., in Short category the gain of 28 bytes - 38.7% and in Simple category the gain is 97 bytes - 54%) in message size,
introduction of YASchema adds more bytes to the total message size. However, YASchema is always advantageous when there is a need to send multiple messages with the same structure (as transmission of the schema will be only a one-time payload).

5.3 Transmission Speed

Three classes of record sets with varying levels of complexity are formed for this analysis, which include Simple, Medium, and Complex. The bytes required for the representation of record sets are considered to represent for the transmission speed. The number of bytes needed to represent each of these record sets is tabulated in Table 4. No packet losses and maximum link speeds (64kbps/256kbps) are assumed for the analysis. It can be seen from Table 4 that all data sizes, especially the Complex type, fit well in the limits of a single TCP/IP maximum segment size when it is in optimized message format.

The time required for transmitting the record sets over the 64kbps link are tabulated in Table 5. The advantages of YAML, especially for the optimized message format, are evident from Table 5. In the case of Medium category, there is an advantage of 15.26 micro secs over XML and 20.09 micro secs over SOAP. The gap is widened for semi optimized YAML SA-YAML (output after stage IV) is getting the maximum advantage (Table 6) by a factor of 2.18 for Small type, 1.71 for Medium type, and 5.14 for Complex type. Similar results can be derived for SOAP record sets also. Thus, the performance of the optimized message format is better than XML and SOAP in all categories. The optimized message format has notable gains over YAML natural format also (6.99, 7.32, and 125.36 micro secs for Small, Medium, and Complex categories, respectively, as can be seen from Table 5).

6 Conclusions

This paper proposes an optimized messaging format for wireless mobile devices based on YAML, a user friendly and lightweight messaging format, supported by YASchema. This formatting requires five stages of processing, starting with YAML. In terms of schema size, message size and transmission speed, this format shows better performance compared to that of XML format. The optimized message format performs better than YAML natural format also. SOAP is also used as a benchmark format, in addition to XML. It is worth mentioning that the proposed method retains the readability of the original message. The proposed optimized message format could be a candidate for standardizing the formats for wireless mobile messaging as it is simple, human readable, easily editable and, most importantly, with less bandwidth requirements.

The future work is to develop this Schema Attached YAML into a messaging standard. This requires many improvements including (i) facility for better encoding in to Abstract Syntax Notation One (ASN.1), (ii) security standards at par with the specifications of XML encryption and related specifications, and (iii) introduction of a policy based security mechanism. The possibility of weaving semantics to the messaging standard is also worth investigating.

References

[10] WAP Binary XML Content Format, Available at http://www.w3.org/TR/wbxm/.
### TABLE 1
Schema Size Analysis

<table>
<thead>
<tr>
<th>Message Categories</th>
<th>Schema Size (in bytes)</th>
<th>Gain %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kwalify (a)</td>
<td>YASchema (b)</td>
</tr>
<tr>
<td>Short</td>
<td>366</td>
<td>121</td>
</tr>
<tr>
<td>Small</td>
<td>987</td>
<td>448</td>
</tr>
<tr>
<td>Medium</td>
<td>1177</td>
<td>576</td>
</tr>
<tr>
<td>Large</td>
<td>1922</td>
<td>899</td>
</tr>
<tr>
<td>Composite</td>
<td>1988</td>
<td>899</td>
</tr>
</tbody>
</table>

### TABLE 2
Comparison - Message Size

<table>
<thead>
<tr>
<th>Type</th>
<th>YAML Original</th>
<th>YAML Phase I</th>
<th>YAML Phase II</th>
<th>XML Original</th>
<th>XML Phase I</th>
<th>XML Phase II</th>
<th>SOAP Original</th>
<th>SOAP Phase I</th>
<th>SOAP Phase II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>145</td>
<td>120</td>
<td>117</td>
<td>221</td>
<td>191</td>
<td>312</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>312</td>
<td>245</td>
<td>215</td>
<td>250</td>
<td>465</td>
<td>604</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>8128</td>
<td>4423</td>
<td>3727</td>
<td>576</td>
<td>10944</td>
<td>13317</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>13799</td>
<td>5901</td>
<td>5872</td>
<td>899</td>
<td>18823</td>
<td>19913</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td>54292</td>
<td>29592</td>
<td>28642</td>
<td>899</td>
<td>73569</td>
<td>74944</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1. System Architecture - Stage V
### TABLE 3
Comparison - Gain Achieved in Message Size (in bytes) with SA-YAML

<table>
<thead>
<tr>
<th>Message Category</th>
<th>YAML Phase II</th>
<th>XML (Excluding Schema)</th>
<th>SOAP (Excluding Schema)</th>
<th>Gain Over (Excluding Schema)</th>
<th>XML (Including Schema)</th>
<th>SOAP (Including Schema)</th>
<th>Gain Over (Including Schema)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
<td>(d)</td>
<td>(c-a)</td>
<td>(d-a)</td>
<td>(c-a+b)</td>
</tr>
<tr>
<td>Short</td>
<td>117</td>
<td>221</td>
<td>191</td>
<td>312</td>
<td>195</td>
<td>-147</td>
<td>-26</td>
</tr>
<tr>
<td>Small</td>
<td>215</td>
<td>250</td>
<td>465</td>
<td>604</td>
<td>250</td>
<td>389</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>3727</td>
<td>576</td>
<td>10944</td>
<td>13317</td>
<td>7217</td>
<td>7445</td>
<td>6641</td>
</tr>
<tr>
<td>Large</td>
<td>5872</td>
<td>899</td>
<td>18823</td>
<td>19913</td>
<td>12951</td>
<td>14041</td>
<td>12052</td>
</tr>
<tr>
<td>Composite</td>
<td>28642</td>
<td>899</td>
<td>73569</td>
<td>74944</td>
<td>44927</td>
<td>46302</td>
<td>44028</td>
</tr>
</tbody>
</table>

### TABLE 4
Record Sets - Categories

<table>
<thead>
<tr>
<th>Type</th>
<th>YAML Phase I</th>
<th>XML Phase I</th>
<th>SOAP Phase I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>XML</td>
<td>SOAP</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>145</td>
<td>120</td>
<td>117</td>
</tr>
<tr>
<td>Medium</td>
<td>320</td>
<td>281</td>
<td>260</td>
</tr>
<tr>
<td>Complex</td>
<td>2171</td>
<td>1182</td>
<td>1144</td>
</tr>
</tbody>
</table>

### TABLE 5
Transmission Speed (in micro secs) using 64kbps link

<table>
<thead>
<tr>
<th>Type</th>
<th>YAML Phase I</th>
<th>XML Phase I</th>
<th>SOAP Phase I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>XML</td>
<td>SOAP</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>17.7</td>
<td>11</td>
<td>10.71</td>
</tr>
<tr>
<td>Medium</td>
<td>39.06</td>
<td>34.3</td>
<td>31.74</td>
</tr>
<tr>
<td>Complex</td>
<td>265.01</td>
<td>144.23</td>
<td>139.65</td>
</tr>
</tbody>
</table>

### TABLE 6
Comparison (SA-YAML Vs XML & SOAP) - Transmission Speed (in micro secs) using 64kbps link

<table>
<thead>
<tr>
<th>Record Set</th>
<th>YAML Phase II (a)</th>
<th>XML Phase II (b)</th>
<th>SOAP Phase II (c)</th>
<th>Gain Over XML (b/a)</th>
<th>SOAP (c/a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>10.71</td>
<td>23.32</td>
<td>38.09</td>
<td>2.18</td>
<td>3.56</td>
</tr>
<tr>
<td>Medium</td>
<td>31.74</td>
<td>54.32</td>
<td>59.45</td>
<td>1.71</td>
<td>1.87</td>
</tr>
<tr>
<td>Complex</td>
<td>139.65</td>
<td>717.9</td>
<td>762.9</td>
<td>5.14</td>
<td>5.46</td>
</tr>
</tbody>
</table>
Fig. 2. Comparison of Message Size (Excluding Schema Size) - SA-YAML Vs XML & SOAP

Fig. 3. Comparison of Message Size (Including Schema Size) - SA-YAML Vs XML & SOAP
Abstract: Black-box and white-box testing are the two major techniques for unit testing. In black-box testing, no information about the internal structure of the program under testing is available. However, in white-box testing, a complete source code or the internal structure is available. Basis path testing is a white-box testing technique that uses a control flow graph (CFG) of a given program to generate a basis set of independent paths for the CFG. Different techniques have been proposed to generate test data that cover all the paths of a basis set. In this paper, we present an interactive tool that performs three tasks:
- It constructs a control flow graph of a given program based on the pseudocode and information provided by the user;
- It computes a basis set of independent paths of the control flow graph;
- It generates test data using genetic algorithm to exercise all basis paths.
We evaluated the performance of different mutation operators for the genetic algorithm based on the percentage of basis paths covered by the generated test data. Experiments show that the use of two known mutation operators, input value and one-point crossover, provide the best path coverage for the programs tested.

Keywords: unit testing, test case generation, control flow testing, genetic algorithms

1 Introduction

Software testing has two main aspects: generating test data and applying test data adequacy criterion. A test data generation technique is an algorithm that generates test cases, whereas a test data adequacy criterion is a predicate that determines whether the testing process is finished. Several test data adequacy criteria have been proposed, such as control flow-based and data flow-based criteria.

The use of genetic algorithms (GAs) in test data generation has become the focus of several recent research studies. This paper presents an interactive tool for automatic test data generation for basis path testing using a genetic algorithm. Our tool consists of modules for constructing the control flow graph (CFG) of a given program based on the pseudocode provided by the user, computing the basis set of independent test paths in the CFG and lastly generating test data using a GA to exercise all the basis test paths. The genetic algorithm conducts its search by constructing a new test data from previously generated test data that was evaluated as effective test data. In the parent selection process, the GA uses the random selection method. In the mutation process, the GA uses five different mutation operators: mutation of input value, one-point crossover, boundary value, uniform value and non-uniform value. We evaluate the performance of the mutation operators in the genetic algorithm on the basis path coverage criterion.
This paper is organized as follows: Section 2 describes the basis path testing technique. Section 3 describes the principles of GAs. Section 4 describes the proposed GA for test data generation and gives the results of applying different mutation operators in the GA to an example program. Section 5 presents the results of the experiments that are conducted to evaluate the effectiveness of the mutation operators in the GA.

2 The Basis Path Technique

Basis path testing is a white-box testing technique that is widely used during unit testing to test the control flow in the program. Basis path testing uses a control flow graph to depict the logical control flow of program under test. A basis set of independent paths is identified from the CFG of a given program (each path starts with a start node and ends with the stop node) and tested to guarantee that all statements and branches in the program have been executed at least once.

The control flow of a program can be represented by a directed graph with a set of nodes (procedure and predicate nodes) and a set of edges. Each node represents a group of consecutive statements. The edges of the graph are the possible transfers of control flow between the nodes. A path is a finite sequence of nodes connected by edges.

Our tool constructs the CFG directly from the pseudocode of a program entered by the user using standard pseudocode notations [Pressman 2003]. The use of pseudocode ensures that there are no syntax errors in the input program and hence in the construction of the CFG.

The CFG of a given program is built from the basic/prime control flow graph notations, using only two operations: sequencing and nesting [Salloum and Salloum 2006]. All programming constructs can be uniquely represented by the prime control flow graph notations. Our tool implements the following programming constructs: sequence (statements), if-then-else loop, while-do loop, do-while loop and for loop. The Boolean constructs (and, or and not) are also dealt with.

Our tool uses the linked list data structure to store both the predicate and procedure nodes created. The predicate and procedure nodes stored in the linked list aid in the derivation of test cases. As the CFG of a program is not a fully connected graph, so it justifies the use of a linked list. Linked list is updated dynamically as user enters the pseudocode. Also, it allows efficient traversal of data nodes when converting the CFG in the shape of a tree for basis path computation.

It is assumed that the parameters and variables in a method call use the built-in data types. A lookup table stores the definition of variables in the given pseudocode. We follow the strongly typed convention of Java language i.e. all variables must first be declared before they can be used. Lookup table holds the following information: variable type, variable name and variable value (if any). Variable type can be any of the built-in data types in Java: integer, long, float, double, short, byte, boolean and character and string. Variable value, if entered, can be either positive or negative. Our program stores the possible range of values for the built-in data types. This range of values, together with the entered variable value is used to generate the initial test data.
The next step is to calculate the cyclomatic complexity of the resultant CFG. Cyclomatic complexity is a software metric that provides a quantitative measure of the logical complexity of the program. The value computed for cyclomatic complexity provides the number of independent paths in the basis set of a program’s CFG. It also provides us with an answer to the number of paths to test in a given program. We calculate cyclomatic complexity by adding 1 to the number of predicate nodes in the CFG.

Our tool implements an efficient and formal algorithm [Salloum and Salloum 2006] to compute a basis set of independent paths from the CFG. The set of basis paths are linearly independent paths that are sufficient to express any other paths of the CFG. For a given CFG, the algorithm constructs a tree structure by visiting every node and edge of the CFG at least once. The paths of the resulting tree from the root to the leaves represent a basis set of paths for the CFG.

3 The Principles of Genetic Algorithms

Genetic algorithm is a heuristic that mimics the evolution of natural species (survival of the fittest) to search for optimal solutions to a problem. GA generates a sequence of populations by using a selection mechanism, and use crossover and mutation as search mechanisms.

Each individual in the population receives a measure of its fitness in the environment. Reproduction selects individuals with high fitness values in the population. Crossover and mutation of fitter individuals derives a new population in which individuals may be even better fitted to their environment. Crossover involves swapping some genes in two individuals. Mutation introduces slight changes to one or several genes in an individual. The structure of simple GA is given below.

Simple Genetic Algorithm ()
{
    initialize population;
    evaluate fitness of population;
    while termination criterion is not reached
    {
        select solutions for next population;
        perform crossover and mutation;
        evaluate fitness of population;
    }
}

The algorithm uses a single crossover and mutation operator throughout the entire genetic process. Also, the algorithm will iterate until the population has evolved to form a solution to the problem or until a maximum number of iterations have taken place (suggesting that a solution is not going to be found given the resources available).
4 A Genetic Algorithm for Test-Data Generation

This section describes a GA that we have implemented for automatic test data generation, which is guided by the control flow dependencies in the program. The algorithm, proposed by Tonella [2004] searches for test cases that achieve maximum path coverage from the basis set of test paths. In other words, the GA in our tool generates test cases for a method under test until a satisfactory level of path coverage (i.e. 100% of method’s paths from the basis set of paths) is attained.

The GA accepts as input the control flow graph of the program to be tested, the set of basis paths to be covered, the number of input variables and the domain and precision of each input variable. Also, it accepts some other GA parameters: population size, maximum number of generations, and the probability of mutation. The algorithm produces a set of test cases, the set of basis paths covered by each test case and the list of uncovered basis paths, if any.

The algorithm uses an integer vector, called the path coverage vector, to record the traversed basis paths. In this vector, each element (initially zero) corresponds to a basis path. Whenever a basis path is covered, the number of the test case that caused this coverage is stored in the corresponding element of the path coverage vector. Our tool keeps track of all the generated test cases that cover new paths in the basis set. We use a counter to count them. These test cases are stored for later use.

Firstly, the overall algorithm is presented, and then the major components of this GA are discussed in turn.

Input:
- Control flow graph of the program P to be tested;
- Basis set of paths to be covered;
- Cyclomatic complexity of the program P;
- Number of program input variables;
- Domain and precision of input data;
- Population size;
- Maximum number of generations;
- Probability of mutation;

Output:
- Set of test cases for program P and set of basis paths covered by each test case;
- List of uncovered basis paths, if any;
**TestCaseGeneration (Input, Output)**

Begin

*Step1: Initialization*

Initialize the path coverage vector to zeros;
Create initial_population randomly based on number of variables in P;
Current_population = initial_population;
Target_paths_to_cover = cyclomatic_complexity of P;
Set of test cases for P = 0;
Coverage_percent = 0;
No_of_generations = 0;
Counter = 0;

*Step 2: Generate test cases*

For each member of Current_population do
Begin
Execute test cases with Current_population as input;
Evaluate the fitness of test case in Current_population;
If (some basis set of paths are covered) then
   Counter = Counter + 1;
   Add effective test case to set of test cases for P;
   Update the path coverage vector;
   Update Coverage_percent;
   Target_paths_to_cover--;
Endif
Endfor
While (Target_paths_to_cover != 0 and Coverage_percent != 100 and
No_of_generations <= Max_gen) do
Begin
   Select set of parents of new population from effective members of
   Current_population according to fitness of test cases using random
   selection method;
   Set of parents of new population = Current_population;
   Create New_population using crossover and mutation operators;
   Current_population = New_population;
   For each member of Current_population do
   Begin
      Execute test cases with Current_population as input;
      Evaluate the fitness of test case in Current_population;
   Endfor
Endwhile
If (some basis set of paths are covered) then
    Counter = Counter + 1;
    Add effective test cases to set of test cases for P;
    Update the path coverage vector;
    Update Coverage_percent;
    Target_paths_to_cover--;
    Break from for loop;
Endif
Endfor
Increment No_of_generations;
Endwhile

Step 3: Produce output
    Return set of test cases for P and set of basis paths covered by each test
case;
    Report the uncovered basis paths, if any;
End.

4.1 Initial population

We randomly generate pop_size to represent the initial population, where pop_size is the
population size. The appropriate value of pop_size is experimentally determined. Initial
population of test cases are generated randomly using the random number generator of
Java language.

4.2. Evaluation function

The fitness of a test case is obtained from the control flow graph edges that are traversed
during its execution. The algorithm evaluates each test case by traversing the CFG, and
checking whether the path covered after reaching the stop node is in the set of basis
paths. If above is true, then the fitness value for the particular test case is closer to one.
Otherwise, the fitness value is close to zero. The fitness value is the only feedback from
the problem for the GA. A test case is considered effective if its fitness value is one. Each
time a target basis path is covered, the test case covering is added to the result set as one
of those necessary test cases to achieve the final level of coverage.

4.3 Selection

After computing the fitness of each test case in the current population, the algorithm
selects test cases from the effective members of the current population that will be
parents of the new population. If none of the members of the current population was effective, all the members of current population are considered the parents of the new population. The selection process of the GA uses the random selection method described below.

In the random selection method, the selection of parents is made randomly, so that every effective member of the current population has an equal chance of being selected for recombination.

Assume that \( n \) members of the current population were effective, where \( n \leq pop\_size \). The parents are selected as follows:

Isolate the effective members and number them from 1 to \( n \);
For \( i = 1 \) to \( pop\_size \) do
  Begin
    Generate a random integer number \( j \) from the range \([0…n]\);
    Select test case \( j \) from the effective members;
  Endfor

4.4 Recombination

In the recombination phase, we use the mutation operator, which is critical to the success of GAs, since it diversifies the search directions and avoids convergence to local optima. The mutation operator creates new individuals from the selected parents to form a new population.

Mutation is used to maintain genetic diversity from one generation of population to the next. The performance of GA is influenced by the choice of mutation operator. Determining which mutation operator to use is quite difficult and is usually learned through experience or by trial-and-error. That is, experiments must be done using all candidate mutation operators to find the best operator for a specific problem, which consumes considerable time and computation resources.

We have used the following known mutation operators in the above genetic algorithm to introduce a small and random change in the next population:

- **Mutation of input value**: involves replacing a value with another random value of the same type.
- **One point crossover**: involves taking the midpoint of current and new population.
- **Boundary value**: involves replacing the chosen value with either the upper or lower bound (chosen randomly).
- **Uniform value**: involves replacing the chosen value with a uniform random value selected between upper and lower bound.
- **Non-uniform value**: involves replacing the chosen value with a non-uniform random value selected between upper and lower bound.
In the traditional GAs, the population evolves until one individual, which represents the solution, is found. In our case, this would correspond to many groups of test cases and several runs of the program to achieve maximum path coverage of the given program (i.e. traversing all the basis paths of a given program). We record which basis paths of the program each test case has covered and halt the evolution when a set of test cases has traversed the entire basis paths of the program. The solution is this set.

5 Experimental Results

The algorithm presented above gives acceptable results for constructing the control flow graph from the pseudocode of a given program. Some of the programs used to test CFG construction include bubble sort, insertion sort, selection sort, quick sort’s partition algorithm and sum of squares from Standish [1998]. It has been observed that CFG for large programs (with four or more predicate nodes) is generated in a reasonable time. The CFG updates dynamically and shows all the pertinent control flow information. The algorithm implemented in the tool for generating a basis set of paths for a given CFG by using Salloum and Salloum’s [2006] approach yields results in linear time.

We evaluated the performance of each of the five mutation operators individually in the genetic algorithm on the maximum basis path coverage criterion. Sometimes, a single mutation operator does not achieve the best mutation results; instead, several mutation operators may have to be applied to achieve the best performance. This observation led us to test the performance of any two mutation operators (out of the five operators) applied simultaneously on the genetic algorithm’s maximum path coverage criterion. The set of mutation operators that provides the best results for path coverage criterion are mutation of input value and one point crossover when applied simultaneously.

Figure 1 shows a table comparing the performance of the five mutation operators applied individually and in combinations of two. The table gives a list of test programs along with the number of paths in the basis set of each program. These test programs were run on our tool. The mutation operators are numbered 1 to 5 in the table where 1 is mutation of input value, 2 is one point crossover, 3 is boundary value, 4 is uniform value and 5 is non-uniform value.
Results from the experiments show that maximum (100%) level of path coverage is achieved when mutation of input value and one-point crossover mutation operators are applied simultaneously. The genetic algorithm implemented in our tool for generating test data achieves optimal coverage of a method’s branches within a reasonable computation time. The resulting test suites are generally compact.
6 Conclusions

Unit testing makes heavy use of white-box testing techniques, specifically basis path testing. Basis path testing uses the control flow graph of a program to generate a set of independent control flow paths.

Our work focuses on generating test cases for basis path testing using genetic algorithms. The GA technique presented in this paper is guided by control flow dependencies in the given program to search for test data that fulfills the maximum path coverage criterion.

The basic steps followed in our work are:

a) Control flow graph construction
b) Basis set of paths selection
c) Test case generation using genetic algorithms

Experiments on our tool show that the algorithms implemented for control flow graph construction, basis set of paths selection, and test case generation for a given program yield results in a reasonable time. The CFG updates dynamically and shows all the pertinent control flow information. Furthermore, the use of genetic algorithms for basis path testing proves extremely powerful. The simultaneous use of two known mutation operators, mutation of input value and one point crossover, achieves 100% basis path coverage.

The use of genetic algorithm with basis path testing has shown satisfying results in terms of path coverage, time taken to generate test cases and compactness of the resulting test suites. A software tool has been implemented and the power of the approach demonstrated.
References


Integrating Generative and Aspect-Oriented Technologies For Framework Instantiation

R. P. Mahapatra
Asst. Professor,
S.R.M. University, Department of Computer Science and Engineering)Chennai.
email : mahapatra.rp@gmail.com

Dr. Jayaram Pradhan
Professor
Berhampur University,
Department of Computer Science, Orissa
email : jayarampradhan@hotmail.com

Abstract
Application frameworks have been successfully used as valuable tools to improve software quality while reducing development efforts. Nevertheless, frameworks still face important challenges in order to be widely adopted. In particular, framework instantiation is still a painful task requiring application developers to understand the intricate details surrounding the framework design. Some approaches to alleviate this problem have already been proposed in the literature but they are usually either just a textual cross-referenced document of the instantiation activities or too tied to technology or specific application domains. In this paper, we present the results of latest investigations to improving our approach to framework instantiation. In particular, we discuss a process language that have been developed to guide framework instantiation explicitly, and the most recent updates that have made to improve the language expressive-ness. Furthermore, we present a case study used to evaluate our approach and to identify current and future extensions.

Keywords: Frameworks, instantiation, software process, software, design, transformation, crosscutting concerns, aspect oriented, generative approach.

1. Introduction
Over the last years, generative programming and aspect-oriented software development have been proposed aiming at increasing maintainability and reusability of software systems. While several research works have focused on the investigation of the individual use of each of these software engineering approaches, less attention has been paid to the integration of these two techniques. Aspect-Oriented Software Development (AOSD) is an evolving approach to modularize crosscutting concern that existing paradigms (e.g.:
object-oriented) are not able to capture explicitly. Crosscutting concerns are concerns that often crosscut several modules in a software system. AOSD encourages modular descriptions of complex software by providing support for cleanly separating the basic system functionality from its crosscutting concerns. Aspect is the abstraction used to modularize the crosscutting concerns. The use of aspect-oriented techniques in the definition of a generative approach can bring additional benefits for the development of system families, such as: (i) clear separate-ion of orthogonal and crosscutting features in the problem and solution space; and (ii) direct mapping of crosscutting features in aspects. Despite these advantages, we believe that the integration of GP and AOSD techniques is not a trivial task. Interest-ing questions arise and need to be considered when developing an aspect-oriented generative approach, including How to model crosscutting features in the problem space? How to design aspect-oriented architectures that address the crosscutting and non-crosscutting features modeled? Which technologies (domain-specific langu-ages, frameworks) are appropriate to implement these aspect-oriented generative approaches for framework instantiation?

Framework concepts have been successfully employed as important tools to achieve software re-use. At the same time they reduce developments efforts and increase the overall quality of produced software systems. ET++, MacApp, Hotdraw, MFC, just to name a few, are important examples of early frameworks that helped demonstrate the feasibility of a framework-centered development approach. They were able to capture common features successfully and represent the variability of a family of applications within a specific domain. There are now a large number of frameworks that have been developed for a variety of different purposes including CORBA (middleware for distributed systems), JADE (agent systems), Strut (web applications), JBoss-AS (enterprise applications), and JUnit (application testing). However, as frameworks become popular their weaknesses as well as their strengths are becoming apparent. In particular, framework instantiation is still a painful task because application developers must understand the intricate details surrounding the framework design. Thus, instantiation

Framework concepts have been successfully employed as important tools to achieve software re-use. At the same time they reduce developments efforts and increase the overall quality of produced software systems. ET++, MacApp, Hotdraw, MFC, just to name a few, are important examples of early frameworks that helped demonstrate the feasibility of a framework-centered development approach. They were able to capture common features successfully and represent the variability of a family of applications within a specific domain. There are now a large number of frameworks that have been developed for a variety of different purposes including CORBA (middleware for distributed systems), JADE (agent systems), Strut (web applications), JBoss-AS (enterprise applications), and JUnit (application testing). However, as frameworks become popular their weaknesses as well as their strengths are becoming apparent. In particular, framework instantiation is still a painful task because application developers must understand the intricate details surrounding the framework design. Thus, instantiation
of a specific application can often be a slow and costly process. For some frameworks such as the MFC it may take up to 12 months for an application developer to be highly productive. Thus, the instantiation process is a time-consuming activity which is counter to one of the most valuable properties of reuse, i.e., significant shortening in development time. Some approaches to alleviate the framework instantiation problems we mentioned have been proposed in the literature. However, they are normally either just a textual cross-referenced documentation of the instantiation activities or too closely to technology or specific application domains.

In this paper we initially give an overview of an approach to framework instantiation. We present RDL (Reuse Definition Language) a process language created to represent framework instantiation activities explicitly. RDL along with xFIT, our supporting instantiation tool, operates on UML models through transformations in order to produce valid application instances. Finally, we discuss a case study we conducted to assess how our framework improved framework instantiation.

The sections of this paper are organized as follows. Section 2 presents our approach to framework instantiation and its latest enhancements. Section 3 depicts the case study we carried out in order to assess our approach properly. In section 4 we discuss how our approach improved framework instantiation in the light of our case study evidences. Section 5 includes some related work, and finally Section 6 presents our conclusions and future work.

2. Understanding Framework Instantiation Problems

However, developers find there is still a steep learning curve when extracting the design rationale and understanding the framework documentation during framework instantiation. Thus, instantiation is a costly process in terms of time, people, and other resources. The framework specifications are usually unstructured or loosely structured and use natural language to describe the artifacts and processes. Even when frameworks have reasonably clear and detailed documentation and modeling. Abstractions (e.g., UML) are used in the design, maintenance today is still performed in reality using source code. Instantiation tasks are not explicit, but remain as tacit.
assumptions in the developers’ minds. In this context, there is a need for extended design abstractions that can allow framework instantiation to be explicitly represented. As a result, framework instantiation is currently a time-consuming activity, which negates one of the most valuable properties of reuse, that is a significant reduction in application development time. In fact, achieving high productivity with framework reuse can take months of effort. Another problem relates to the consistency of the final application design. Some instantiation processes introduce unexpected states that can violate some of the framework design constraints. Therefore, it is important that the framework documentation provides a set of properties (structural and behavioural) that must be preserved after each instantiation process is performed. In summary, design modelling notations need to be extended and the instantiation tasks explicitly represented at the design level and, in addition, there is a need for methods that allow validation of these tasks.

3. An Aspect-Oriented Generative Approach

The aspect-oriented (AO) generative approach aims at exploring the horizontal domain of multi-agent systems (MASs) to improve their quality and productivity. The purpose of the generative approach is threefold: (i) to uniformly support cross-cutting and orthogonal (noncrosscutting) features of software agents starting at early development stages; (ii) to abstract the common and variable features; and (iii) to enable the code generation of AO agent architectures.

Figure 1 depicts our generative approach that is composed of:

(i) a domain-specific language (DSL), called Agent-DSL, used to collect and model orthogonal and crosscutting features of software agents;
(ii) an AO architecture modeling a family of software agents. It is centered on the definition of aspectual components to modularize the crosscutting agent features;
(iii) a code generator that maps abstractions of the Agent-DSL to
specific compositions of objects and aspects in agent architectures.
The definition of our generative approach encompassed a typical domain engineering process. The steps followed in the development of the generative approach were:

1. Domain Analysis
   a. Definition of the domain
   b. Identification and modeling of common and variable features of the domain
   c. Identification and modeling of the crosscutting features of the domain

2. Domain Design
   a. Specification of the generic AO architecture
   b. Identification and specification of the DSLs
   c. Specification of the configuration knowledge

3. Domain Implementation
   a. Implementation of the DSLs
   b. Implementation of the AO architecture and additional components
   c. Implementation of the code generator

4. Our Approach to Framework Instantiation

A typical framework adaptation has two phases: i) understanding the overall rationale behind the framework design; ii) extending the framework flexible points according to specific requirements in order to produce application specific increments (ASI).

As we have mentioned, the first phase has been supported by some framework documentation approaches. Basically, they describe the purpose of the framework, its major design elements, their relationships, how the flexible points can be adapted to produce applications, and provide some examples. For example, in the cookbook approach, recipes are used to explain how a certain extension point can be adapted. Recipes can reference each other, thus helping application developers to understand better how the hotspots (and the design elements they represent) are interrelated.

Our approach complements framework documentation techniques, in particular cookbooks. It closes the gap left by purely text-based approaches by providing means to represent instantiation activities explicitly. Our approach consists of a process language, RDL, that allow framework developers to represent adaptation steps, and a supporting tool, xFIT, that operates on UML models by transforming a framework’s class diagrams into application class diagrams based on application
developer’s inputs. Figure 1 below depicts our approach.

Figure 1: Overview of Our Approach

The steps required to instantiate a framework using our approach consists of:

- The framework developer provides a framework class diagram conforming with the XMI format (an XML file representing the model);
- The framework developer provides an RDL script containing the framework instantiations steps;
- The application developer runs xFIT providing it with the RDL script and the framework UML class diagram. Likewise, the application developer provides feedback according to specific application requirements.

- At the end of the generation process xFIT will run validation tasks and report the errors encountered (if some exist). Otherwise, a UML class diagram is produced including the framework and the specific application instance classes.
- The application developer can then use a Case tool to open the application model and generate stubs for the classes produced. By filling out the stubs with appropriate code the process is ended.

4.1 The Process Language (RDL)

RDL is a process language that aims at providing mechanisms for framework developers to represent instantiation tasks explicitly. RDL is programming-language and framework-domain independent and manipulates design elements expressed in UML. RDL abstractions have been proposed based on the cookbook approach and exploit the use of design patterns.

In the next sections we describe the main construct of the RDL and the latest enhancements we have made to increase the language expressiveness. Since it is not the purpose of this paper
to be an RDL reference manual we suggest reading for a more detailed description of the language.

**RDL Main Structure** - RDL higher level constructs are represented by cookbooks, recipes and patterns. AN RDL cookbook contains a set of RDL recipes. RDL recipes embody instantiation tasks related to particular variable aspect of a framework architecture. RDL Patterns describe recurring instantiation steps encountered during a framework adaptation (e.g. design patterns).

RDL can be used to produce two types of artefacts: RDL scripts and Pattern Libraries. A general structure of an RDL script is shown in Table 1.

```
COOKBOOK myCookBook
RECIPE main
...
CALL_RECIPE( R1, (…) );
...
END_RECIPE
RECIPE R1(…)
...
END_RECIPE
...
END_COOKBOOK
```

Table 1: General structure-RDL Script

In an RDL script, at least one recipe must be named `main` representing the cookbook start point. Recipes can call each other, receive parameters and return values in a way similar to functions in procedural languages.

RDL Pattern Libraries describe instantiation patterns, i.e., recurring instantiation tasks. Since some design patterns exhibit an abstract and a concrete part (e.g. Template Method, Abstract Factory, and Strategy) they can be properly used to expose framework hotspots. Therefore, design pattern instances can be represented as RDL Patterns. Pattern Libraries represent an enhancement we have made to our approach. The general structure of an RDL Pattern Library can be found in Table 2.

```
PATTERN_LIBRARY
myPatternLibraryName
PATTERN Pattern1(…)  
...
END_PATTERN
PATTERN Pattern2(…)  
...
END_PATTERN
...
END_PATTERN_LIBRARY
```

Table 2: General structure-RDL Pattern Library

Pattern libraries are normally stored in files with the `.rdp` extension (reuse definition pattern). UML class models are expected to conform to the XMI
format and are stored in .xmi files. In the current version of RDL only one RDL script (.rdl) is allowed to specify the instantiation steps of a framework. There is still no way to import and combine RDL scripts. This has been left as a possible enhancement for future versions of our language.

**RDL Types** - In order to keep the syntax of the language simple, the previous versions of RDL did not consider data types explicitly. However, as we used the language in practical situations the need for a strong typed-language became apparent. Thus, the types now encountered in RDL are basically those found in UML class diagrams plus some additional ones to represent strings, numbers, booleans and list of types (Table 3). Each type has a set of associated operations and attributes that allow framework developers (RDL script users) to make proper references to model elements.

<table>
<thead>
<tr>
<th>RDL Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRING</td>
<td>Represent Strings in RDL</td>
</tr>
</tbody>
</table>

RDL commands fall into three categories: Basic, Instantiation, and Pattern Commands. Following we discuss each one of the categories.

**RDL Basic Commands** - The basic commands provide low-level facilities to manipulate the framework design elements. For instance, new classes, methods or attributes can be created and added to UML class diagram models. Table 4 below illustrates some of the RDL basic commands.

<table>
<thead>
<tr>
<th>Description</th>
<th>Basic Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class creation</td>
<td>NEW_CLASS(…)</td>
</tr>
<tr>
<td>Method creation</td>
<td>NEW_METHOD(…)</td>
</tr>
<tr>
<td>Attribute creation</td>
<td>NEW_ATTRIBUTE(…)</td>
</tr>
<tr>
<td>Inheritance</td>
<td>NEW_INHERITANCE(…)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RDL Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER</td>
<td>Represent Numbers in RDL</td>
</tr>
<tr>
<td>BOOLEAN</td>
<td>Represent Booleans in RDL</td>
</tr>
<tr>
<td>PACKAGE</td>
<td>Represent UML Packages</td>
</tr>
<tr>
<td>CLASS</td>
<td>Represent UML Classes</td>
</tr>
<tr>
<td>METHOD</td>
<td>Represent UML Class Methods</td>
</tr>
<tr>
<td>ATTRIBUTE</td>
<td>Represent UML Class Attributes</td>
</tr>
<tr>
<td>Lists</td>
<td>Represent lists (vectors) in RDL</td>
</tr>
</tbody>
</table>

Table 3: RDL Types
Instantiation Commands- RDL

Instantiation Commands increase the level of abstraction by combining basic commands into single tasks. Basically, RDL Instantiation Commands represent object-oriented reuse activities such as extending a class, overriding a method, and assigning a value to a class attribute. Table 5 depicts the main Instantiation Commands.

<table>
<thead>
<tr>
<th>Description</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Extension</td>
<td>CLASS_EXTENSION(…;)</td>
</tr>
<tr>
<td>Method Extension</td>
<td>METHOD_EXTENSION(…;)</td>
</tr>
<tr>
<td>Value Assignment</td>
<td>VALUE_ASSIGNMENT(…;)</td>
</tr>
<tr>
<td>Value Selection</td>
<td>VALUE_SELECTION(…;)</td>
</tr>
</tbody>
</table>

Table 5: Main RDL Instantiation Commands

Pattern Commands- The highest level statements in RDL are represented by Pattern Commands. Pattern Commands allow framework developers to reuse a set of recurring instantiation activities previously specified. In previous versions of RDL, Patterns Commands were represented by the Pattern Class Extension and Pattern Method Extension commands. These commands required specific types to be passed as input parameters in order to be used properly. We decided to simplify the language support for patterns by defining a single command for an RDL Pattern call. No parameters are required and it is up to the framework developer to define how patterns will be properly described. Table 6 describes the RDL command to call an RDL Pattern.

<table>
<thead>
<tr>
<th>Description</th>
<th>Pattern Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern call command</td>
<td>CALL_PATTERN(…;)</td>
</tr>
</tbody>
</table>

Table 6: RDL Pattern Call Command

In the following we illustrate the implementation of an RDL Pattern Library including an implementation for the Factory Method design pattern, and an RDL Script benefiting from the RDL Pattern implementation.

PATTERN_LIBRARY

GammaPatterns

PATTERN FactoryMethod( IN absCreatorName : STRING , IN facMethodName : STRING , INOUT concreteCreatorClass : CLASS )

-- Create Concrete Creator

67
IF (concreteCreatorClass = NIL)
concreteCreatorClass =
CLASS_EXTENSION(
absCreatorName, ? );
END_IF
-- Extends Creator Factory Method
METHOD_EXTENSION(
absCreatorName, facMethodName,
concreteCreatorClass );
END_PATTERN
END_PATTERN_LIBRARY
COOKBOOK myCookBook
RECIPE main
conCreator : CLASS;
CALL_PATTERN(
FactoryMethod, ( 
“AbstractView”,
”createAlarm”, conCreator));
END_RECIPE
END_COOKBOOK

4.2 The Framework Instantiation Tool (xFIT)
Our approach is supported by an instantiation tool known as xFIT (Framework Instantiation Tool). xFIT provides a runtime environment for RDL scripts. The framework class diagram and an RDL script are taken as inputs and based on application developers feedback xFIT generates the application instance class diagram. xFIT performs validation tasks over the design elements produced to ensure that its structure is regular and well-formed. As an example, xFIT certifies that all abstract classes and methods have been resolved in the final design since all the hotspots must have been handled.

5. CASE STUDY
We developed an initial case study to evaluate the feasibility of our approach. The idea was to produce an application instance through transformations over an AO framework design. We developed a Drawing Editor framework in AspectJ which exposed 6 hotspots (CSG Drawing Editor Framework). The Observer design pattern was used allowing Figure objects (Subjects) to notify registered Observers about size changes (e.g., zoom in/out). The Observer, in our case represented by the Display class would handle the notifying events by repainting the drawing in the appropriate canvas window. We used an aspect version of the Observer design pattern that exposes 2 hotspots i) the Display (Observer) reaction to Figure’s (Subjects) resizing, realized by an aspect method extension, and ii)
the Figure object’s that should notify the Display about state changes (e.g., calls to the resize() method of Figure objects), realized by a pointcut extension. We represented the instantiation tasks in an AF-RDL script, mapped some AF-RDL commands to XQuery user-defined functions and performed the corresponding XQuery transformations. At the end we obtained a serialized XMI-like model representing our application instance design. Aspects, pointcuts, aspect methods and advices were represented by specific model elements. The case study showed that the idea of mapping high-level AF-RDL commands to XQuery functions is feasible. Indeed, XQuery turned out to be a powerful language for transformation. Therefore, the combination of AF-RDL and XQuery was seen as very positive and promising. The next steps in our research include i) enhancing the AF-UML expressiveness by defining new models for AO frameworks, ii) specifying a corresponding XMI-compliant description to all model elements in AF-UML, iii) mapping all RDL instantiation commands to related XQuery functions, and iv) developing a tool to manipulate AF-UML models.

6. Conclusion

This paper reported our experience in the definition of an AO generative approach. The goal of this approach is to explore the horizontal domain that MASs represent in order to enable the code generation of agent architectures. During the development process of the generative approach, it was necessary to adapt modeling notations used in generative programming due to the adoption of AOSD. Also, a new notation was proposed to support the representation of AO architectures (section 3). Aspectsual components have been used to model crosscutting features from the architectural point of view. We believe that the definition of AO generative approaches can bring important benefits to the development of software families. GP allows: (i) to evolve the problem and solution spaces independently; and (ii) to define clearly the mapping between high-level features and implementation components. The integrated use of GP and AOSD techniques brings additional benefits, such as: (i) clear separation of orthogonal and crosscutting features starting at early design phases; and (ii)
direct mapping of crosscutting features in aspectual components. This latter benefit simplifies the implementation of code generators, because the composition of crosscutting concerns is accomplished by the aspect weavers. Using only OO abstractions, crosscutting agent features need to be hand-coded in the code of classes. This work aimed at identifying relevant techniques and requirements to be considered on the development of AO generative approaches. It represents a significant step in the definition of a method to develop AO generative approaches.

we also presented our approach to framework instantiation, it shows the enhancements that have been made to improve its effectiveness, and a case study to evaluate our approach. Our case study showed that the use of our approach was helpful in raising the level of abstraction of framework adaptation processes.

7. References

Semi-Automatic Generation of Transformation Rules in Model Driven Engineering: The Challenge and First Steps

Slimane Hammoudi, Samyr Vale

ESEO, Ecole Supérieure de l’Ouest
4, Rue Merlet de la Boulaye BP 926,
49009 Angers cedex 01
France
slimane.hammoudi@eseo.fr; samyr.vale@eseo.fr;

ABSTRACT: Recently, Model Driven Engineering (MDE) approaches have been proposed for supporting the development, maintenance and evolution of software systems. Model driven architecture (MDA) from OMG (Object Management Group), “Software Factories” from Microsoft and the Eclipse Modelling Framework (EMF) from IBM are among the most representative MDE approaches. Nowadays, it is well recognized that model transformations are at the heart of these approaches and represent as a consequence one of the most important operations in MDE. However, despite the multitude of model transformation languages proposals emerging from university and industry, these transformations are often created manually, generally a fastidious and error-prone task, and therefore expensive process. We argue that, the semi-automatic generation of transformation rules is an important challenge in future MDE development. In this paper we propose an extended architecture that aims to semi-automate the process of transformation in the context of MDA. This architecture introduces mapping and matching as first class entities in the transformation process, represented by models and metamodels. We will introduce and discuss briefly two main operations “adaptation” and “derivation” which we consider as core techniques for a semi-automatic transformation process in MDA, along with the first two main techniques of matching and mapping. Finally, we will present a classification of the major approaches to matching in the literature and situate them in the context of MDE metamodel matching. These matching techniques are the centerpiece of a semi-automatic transformation process.

KEYWORDS: model driven architecture, transformation language, mapping metamodel, matching metamodel, semi-automatic transformation, transformation architecture and matching techniques.

1. Introduction

The main motivation behind model driven engineering (MDE) is to transfer the focus of work from programming to modeling by treating models as first class entities and consequently the primary artifacts of development. There are nowadays several approaches based on MDE principles, the most well known being MDA [20] by OMG , “Software factories” by Microsoft [10] and the Eclipse Modelling Framework (EMF) from IBM [7]. In the literature, several issues around MDE have been studied and subject of intensive research, e.g. modeling languages [3] [5], model transformation languages [14] [21], mapping between metamodels [11], and design methodologies [1]. Among these issues, model transformation languages occupy a central place and allow to define how a set of elements from a source model are analyzed and transformed into a set of elements of a target model. However, these transformations are created manually, often a fastidious and error-prone task, and therefore an expensive process. These transformations consist of creating a set of rules involving, and in the same time merging mapping and transformation techniques between two metamodels. A semi-automation of the transformation process leads to a real challenge allowing many advantages: It enhances significantly the development time of transformation and decreases the errors that may occur in a manual definition of transformations. In [17], we have initiated a first attempt towards this semi-automation. We have introduced an approach separating mapping specification from transformation definition, and have implemented this approach in a tool called MMT (Mapping Modeling Tool). In this first approach, a mapping specification was created manually to define the relationships between metamodels (i.e. equivalent metamodel elements), while
transformation definition was generated automatically and contained the operational description of the transformation rules between models. In this work, we propose to push the semi-automation process one step further by considering matching techniques [15] [22], to provide semi-automatic mappings between two metamodels. The produced mappings could be adapted and validated by an expert for the automatic derivation of a transformation model, as a set of transformation rules. In this paper, we present an extended architecture of the transformation process in the context of MDA. This architecture introduces the matching and mapping components as two other important operations in the transformation process. We will introduce and discuss briefly two main operations “adaptation” and “derivation” which we consider as core techniques for a semi-automatic transformation process in MDA.

This paper is organized as follows: section 2 introduces the core concepts of model transformation in MDA and point out the main problems of the transformation process. Section 3 presents an extended architecture for a semi-automatic transformation process and discusses the matching and mapping metamodels as two important components in this process. Section 4 reviews the matching techniques that have been proposed in the literature and situates metamodel matching in the context of MDE. Finally, section 5 concludes our work and presents some final remarks and perspectives.

2 Model Transformations: Core concepts and main problems

It is well recognized today that model transformation is one of the most important operations in MDA. In our discussions here we are concerned with a transformation that takes a platform-independent model and transforms it into a platform-specific model. In the context of the basic four levels Metamodeling architecture of MDA [20], various scenarios of model-to-model transformation have been identified. Figure 2 presents the most common scenario of these transformations, which is compatible with the MOF2.0/QVT standard [21]. Each element presented in Figure 2 plays an important role in MDA. In our approach, MOF is the well-established metametamodel used to create metamodels. Transformation rules specify how to generate a target model (i.e. PSM) from a source model (i.e. PIM). To transform a given model into another model, the transformation rules map the source into the target metamodel.

![Figure 1. Model Transformation in MDA: from PIMs to PSMs](image)

The transformation rules are based on the transformation language, such as the standard QVT. The transformation engine takes the source model, executes the transformation rules, and produces the target model as output. We point out two main problems concerning this main scenario of the MDA transformation process illustrated by figure 2:
The first problem concerns the creation of “transformation rules” between metamodels which, as mentioned in the introduction, are often created manually using a transformation language, generally a fastidious and error-prone task, and therefore expensive process [8].

The second problem concerns the specification of these “transformation rules”, which merge together techniques of mappings and transformations without explicit distinction between them. That is to say, the specification of correspondences between elements of two metamodels and the transformation between them are grouped in the same component at the same level. As discussed in [12], an explicit distinction between techniques of mapping and transformation could be very helpful in the whole MDA process of transformation. Moreover, the separation between the mappings and transformations parts is a first step towards a semi-automatic process, since mappings could be discovered and generated by a matching process.

3 An extended architecture for the transformation process

Figure 3 illustrates our proposal of an extended architecture for the transformation process in MDA, allowing a semi-automatic generation of transformation rules between two metamodels, and the semi-automatic generation of a target model from a source model. The first three main operations of our approach are: Matching, Mapping and Transformation. All the components linked to these operations, and their relationships, are presented in figure 3 based on the four level MDA metamodeling architecture.

The matching operation [2] [13] is the process that produces the potential mappings between two metamodels. Generally, this task implies a search of equivalent or similar elements between two metamodels. In the database domain, this task is called schema matching. In our context, a matching model (Matching M) takes two metamodels designed by source and target (representing respectively a PIM and a PSM metamodel), and produces a mapping model (Mapping M). The matching model conforms to a metamodel of matching (Matching MM) which implements techniques that consist of finding semantically equivalent modeling concepts between two metamodels. Thus, different kinds of relationships between metamodel elements are discovered using the metamodel of matching.

The relationships between metamodel elements are saved in a mapping model which conforms to a mapping metamodel (Mapping MM). This metamodel defines the different kinds of links (relationships) that could be generated by the matching model. Each kind of link corresponds to one transformation pattern specified in the transformation model described hereafter. Given that no generic matching solution exists for different metamodels and application domains, it is recommended to give the human expert the possibility to check the obtained mappings, and, if necessary, update or adapt it. This is one of the steps in the whole process, in which the expert intervenes to complete and validate the obtained results.

Finally, a transformation model (Transformation M), in conformance to its transformation metamodel (Transformation MM), is derived automatically from a mapping model. A transformation model is basically represented by a set of rules that states how elements from source metamodel are transformed into elements of target metamodel. These rules are expressed in a transformation language based on MDA standards (OCL, MOF). This language, such as the standard QVT is described by a metamodel as a general formalism and abstract syntax for model transformation in MDA. Frequently, the transformation model is completed by some information such as those concerning the execution environment, and produces a transformation program ready for the execution. This last part is often achieved by a designer (or software engineer) who implements a business model in a specific platform. Finally, a transformation engine takes a source model as input, and executes the transformation program to transform this source model into the target model.
Figure 3. Architecture for a semi-automatic transformation process in MDA.

The first goal with this architecture is to introduce the matching process into the OMG’s MDA approach in order to increase the degree of automation of the transformation process. This leads to a reduction in manual human tasks often fastidious and error-prone, by the rational choice among the plethora of existing works on matching techniques. These techniques are suitable for the problem of automatic mapping production. Thus, from a software point of view, the transformation process involves three main programs which are at the heart of a semi-automatic development:

- **Matching program**: implements the matching metamodel and produces a first version of a mapping model between two metamodels source and target. This mapping model is adapted and validated by an expert user.
- **Generation program**: takes a mapping model validated by an expert, and derives automatically a transformation model (program) as a set of rules.
- **Transformation program**: takes a source model defined by a designers or engineers and produces an equivalent target model on a specific platform.

Two important operations adaptation and derivation allow to link and complete these main programs in the whole process of transformation. Adaptation is the responsibility of the expert user who should accept, discard or modify the obtained mappings, furthermore, to specify the correspondences which the matcher was unable to find. Loosely speaking, the mapping and matching techniques (models) could be defined with the following intuitive formula:

\[ \text{Mapping} = \text{Matching} + \text{Adaptation} \]

The mapping model obtained in the previous step after adaptation by the expert user should be completely defined allowing an automatic generation of transformation model. This operation is called derivation and, in the same way as above, transformation and mapping models can be defined with the following intuitive formula:

\[ \text{Transformation} = \text{Mapping} + \text{Derivation} \]
4 Matching techniques for metamodels in MDE

Matching between metamodels are the centerpieces for a semi-automatic transformation process in MDE and MDA in particular. Matching techniques have been studied in various research domains, including digital libraries, ontologies, agent matchmaking, schema integration and evolution in databases ([13], [22]). In the context of MDE, we can find very few works in the literature that address the problem of metamodels matching. To the best of our knowledge, [15] is the only work that investigate metamodels matching. They propose the application of ontology and schema matching techniques for automatically exploring semantic correspondences between metamodels. Moreover, in this work they introduce a lifting process, which allows to create ontologies from metamodels, and after they apply techniques of matching between ontologies. Schemas in the context of databases and metamodels in our context of MDE are closely related, hence, we propose to review the different approaches of schema matching, and after we situate these approaches in our context of metamodeling matching.

4.1 Classification of schema matching approaches

In the literature, several schema matching approaches have been proposed [13] [22]. Each schema matching approach has its own characteristics that were grouped in a taxonomy illustrated bellow in figure 3 [9] [18]. In addition, each approach has been evaluated through match quality measures discussed in the next section 4.2.

![Figure 3. Classification of schema matching approaches.](image)

- Individual matcher approaches use only one matching criterion. They are classified in:
  - **Schema-only based**, when they consider only metamodels. They can be classified in:
    - **Element level**, the mapping is realized for each individual element. It can be classified in linguistic and constraint-based. Linguistic are based on name similarity, description, global namespace, while constraint-based are based on type similarity and key properties.
    - **Structure-level**, the mapping is realized considering the combinations of elements related in a structure. It is only classified in constraint-based that use graph matching.
  - **Instance/contents-based**, when they consider only instances (or models). It can also be classified in element-level. This last can be classified in linguistic and constraint-based. In
this case, linguistic is based on word frequencies and key terms present in the element instances, while constraint-based is based on value pattern and ranges of the element instances.

- Combining matchers use multiple matching criteria. They can be classified in:
  - **Hybrid**, they combine multiple approaches to create only one matcher in order to produce a result, i.e. the creation of mapping between the elements.
  - **Composite**, they combine many results obtained from different approaches in order to produce the mapping between elements. This combination of results can be manual or automatic.

### 4.2 Matching quality Measure

The interrelationships between metamodels can be organized in sets which can be manually or automatically created. A set created manually can contain all needed matches (i.e. matched elements); while a set created automatically can contain valid and non valid matches. The first set is denominated real matches, and the later derived matches (cf. Figure 4).

![Figure 4. Comparing real matches and automatically derived matches.](image)

In addition, other subsets are defined as follows [13] [18]:

- **A** (false negatives) are matches needed but not automatically identified.
- **B** (true positives) are matches which are needed and have also been correctly matched by the automatic match operation.
- **C** (false positives) are matches falsely proposed by the automatic match operation.
- **D** (true negatives) are false matches which have also been correctly discarded by the automatic match operation.

Based on the cardinalities of these sets, the following match quality measures are provided as parameters for benchmarks:

\[
\text{Precision} = \frac{|B|}{|B| + |C|} \quad \text{reflects the share of real correspondences among all found ones.}
\]

\[
\text{Recall} = \frac{|B|}{|A| + |B|} \quad \text{specifies the share of real correspondences that are found.}
\]

\[
\text{F-Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{Overall} = \text{Recall} \times (1 - \frac{1}{\text{Precision}})
\]

All these measures were developed specifically in the schema matching context [9] [18]. We can notice that F-Measure represents the harmonic mean of Precision and Recall. The main underlying idea of Overall is to quantify the post-match effort needed for adding missed matches and removing false ones.
4.3 From schema matching to metamodel matching

In our MDE context with respect to our extended architecture of figure 3, metamodel matching results in a mapping model that must be conform to a mapping metamodel. In [16] [17], an initial mapping metamodel was proposed and implemented in a tool called MMT. According to model management algebra [2], a mapping is generated using an operator called match which takes two metamodels as input and returns a mapping between them. We have adapted this operator as follows: given two metamodels Ma and Mb, and C Ma → Mb, the mapping model (a set of correspondences) which conforms to the mapping metamodel MC. The operator match could be defined formally as:

\[ \text{Match}(M_a, M_b) = C_{M_a \rightarrow M_b} \land MC. \]

In general, metamodels are created with a specific purpose and by different groups of persons. Each purpose is determined in function of the domain, and each group of persons models a system in different ways. In the modeling task, each group abstracts, classifies and generalizes the reality based on its own knowledge. Consequently, metamodels that were created in the same context and by different groups may have different structure and terminology causing the semantic distance among them [18]. According to our approach, a model can be transformed in another model, only if the metamodel of the former can be mapped in the metamodel of the later. In order to map metamodels, the equivalent or similar elements must be identified, and the semantic distance should be minimized. The notion of semantic distance was developed to cover the notion of “how close is close enough”. A dual for “semantic distance” is schema similarity that is defined as “the ratio between the number of matching elements and the number of all elements from both input schemas” [22]:

\[ SS = \frac{N_m}{N_t}, \]

where SS is the schema similarity, Nm is the number of matching elements and Nt is the number of all elements. Semantic distance can also be quantified as a numeral value (like schema similarity) or as a subset of a metamodel. By the way, according to the MDA manifesto [6], “one of the primary purposes of automation in MDA is to bridge the semantic gap between domain concepts and implementation technology by explicitly modeling both domain and technology choices in frameworks and then exploiting the knowledge built into a particular application framework”. Moreover, automation, which is the main objective of this work in the context of model transformation, is one of the basic tenets of MDA manifesto [6].

5. Conclusion and Future Work

In this paper, we have presented a first approach for a semi automatic transformation process in MDA using an extended architecture. We argue that a semi-automatic transformation process will be a great challenge in MDA as there is not yet a complete solution that automates the development of model transformation. A semi-automatic process will bring many advantages: it accelerates the development time of transformations; it reduces the errors that may occur in manual coding; and it increases the quality of the final transformation code. The key principle for this process is to consider mapping and matching metamodels as first class entities in MDA. In our previous work [18] [19], we have proposed a first algorithm for metamodel matching based on set theory. In future work, we will propose a methodology for a semi-automatic transformation process. This methodology will enforce our architecture and details the different steps in the semi-automatic process, and the main MDA users involved in this process. We will also implement our proposed matching algorithm in the context of our methodology and we will investigate from the obtained mappings, the adaptation and derivation techniques discussed in section 3.
6. Bibliography


Use of Simulation in Systems Engineering

Hassan Badkoobehi, Ph.D.
National University, San Diego

Abstract

As used in systems engineering, the word simulation refers to the construction of a simplified representation of a process or system in another to facilitate its analysis. Such a representation or model may be quantitative or qualitative. In either case it is characterized by the fact that it does not include all the features and characteristics of the original system or process. Rather the purpose of the simulation is to show the effect of particular factors which are being investigated. The four major areas I will discuss in my paper would be:

1. The Development of Models
2. Stochastic and Deterministic Models
3. The Uses of Simulation
4. Simulation Equipment

All the types as described above may be referred to as simulation, and will be so considered in this paper.

In this paper, various aspects of the simulation method have been presented, the general characteristics with analog, digital, and hybrid techniques has been discussed. The importance of simulation in systems engineering arises from the fact that simulation makes possible the verification of proposed designs before completion of system development, thus resulting in invaluable saving in time and money. The major considerations involved in the use of simulation has been discussed in considerable detail. It has been pointed out that simulation is always partial and requires the selection and isolation of significant variables. Consequently the results of a particular simulation are generally applicable only to particular aspects of a system design. Simulation is a tool used in the verification of design hypothesis as well as being a source of new ideas, new designs and new hypothesis. It forms an invaluable link in the process of systems engineering.

The Nature of Simulation

The word simulation refers to the construction of a simplified representation of a process or system in order to facilitate its analysis. Such a representation that simulation does not
include all the features and characteristics of the original system or process, but it shows the effect of particular factors which are being investigated.

**The Development of Models**

The process of constructing a simulation of a physical system thus involves one or more abstractions from the real world. These abstractions may be of varying degrees of severity. For example, the construction of a scale model of the physical system, a procedure common in hydraulics and the chemical industry. While there are important problems of scaling in construction of such a model, in principle or scale model of a harbor, for example represents a relatively small departure from the real world.

**Stochastic and Deterministic Models**

The construction of a model is based on information obtained from the physical world by observation or measurement. Consequently, measurement errors will result in erroneous models. One of the serious problems in the simulation of a process is the selection of those random elements which one desires to incorporate into the model. Many models are constructed on a purely deterministic basis with the understanding that the results obtained from the models may represent statistical averages of certain variables in the physical system.

**The Uses of Simulation**

Simulation in its various forms is of great importance in systems engineering. Some of the important applications are listed below:

1. Design evaluation - a simulated system may be used to evaluate the validity of the preliminary design of a portion of a system.

2. Interrelationships among the parts - one of the key uses of simulation is in the evaluation of the effect of various portions or subsystems of a system upon each other and upon the performance of the system as a whole.

3. Costs - An important part of simulation is in the attempt to reduce overall costs by the evaluation of alternative designs by means of simulation.

4. Study of failure sources - without some type of simulation, it is extremely difficult in many cases to determine possible sources of airplane crashes or controller instability.

5. Hypothesis testing - simulation makes possible experiments which in the physical system may be difficult or impossible to test.

Many other uses and applications of simulation can be listed.
• Determination of problem area
• Determination of significant and insignificant variables
• Study of the effect of environmental variation upon performance.

Simulation Equipment

Simulation equipment may be classified in several ways.
• Analog simulation
• Digital simulation
• Physical simulation
• Mathematical simulation
• Manned system simulation
• The computers and hybrid analog - digital simulation analog simulation

When a simulation is characterized primarily by continuous signals or elements it is referred to as analog simulation. When a substantial portion of a simulation involves discrete signals, elements, or processes, it is referred to as hybrid analog - digital simulation, analog simulations can take both the forms outlined above.

Digital Simulation

When simulation consists of the manipulation of phenomena which occur with discrete values, it is referred to as digital simulation. Digital simulation lends itself to the study of systems wherein there are many decision functions to be implemented, and to phenomena which are more easily characterized by a word description than by a set of conventional mathematical equations. Digital simulation is usually carried out on a general-purpose digital computer. The process of preparing the simulation model for a digital computer follows many of the general rules which arise in programming any problem for the digital computer.

Physical Simulation

Simulation may involve the study of a physical system by means of any analogous system whose behavior closely approximates that the system under study for the particular phenomena being investigated through forms of such simulation are used:

• Scale models are used in hydraulics and in the process industries.
• Analogous models are used in heat or gaseous flow.
• Partial-system tests used in the interconnection of an element of a physical system with a general-purpose computer, which represents a mathematical analog of the system. Mathematical simulation represents a further level of abstraction that physical simulation.

A mathematical simulation of the system to be studied must be formulated. The resulting equations are solved by means of analog computer events which perform specific.
operations such as summation, integration or multiplication. The physical variable in the computer element is generally either an electrical voltage or a shaft rotation. The simulation can be performed using a general-purpose electronic analog computer, or it may be performed using special purpose elements.

**Comparison of Physical and Mathematical Simulation**

<table>
<thead>
<tr>
<th>Mathematical Simulation</th>
<th>Physical Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Easy parameter variations</td>
<td>* Parameter variation may be difficult</td>
</tr>
<tr>
<td>* Mathematical description required of all system elements</td>
<td>* Mathematical description NOT required of all system elements</td>
</tr>
<tr>
<td>* Time scale can be varied by selection of computer components</td>
<td>* Generally designed for a fixed time scale</td>
</tr>
<tr>
<td>* Well suited to fast-time simulation</td>
<td>* Well suited to real-time simulation with human operators</td>
</tr>
<tr>
<td>* Results affected by selection of model and quality of computer components</td>
<td>* Results affected by selection of model and validity of analog</td>
</tr>
<tr>
<td>* Possibility of false solutions due to the characteristics of the equations themselves</td>
<td>* No such possibility</td>
</tr>
</tbody>
</table>

Model for each major system element must be formulated. In cases where such models are not well known or where the performance of particular subsystems is difficult to describe mathematically because of complexity, or where statistical variability of special types is important, it may be necessary to include elements physically in the simulation.

**Manned Systems Simulation**

One of the most important areas of application of the simulation method is in the study of original conceptual design and the final manufacture.

In this paper various aspects of the simulation method have been presented. The general characteristics of simulation have been outlined, and simulation with analog, digital, hybrid, and manned systems has been discussed. Since simulation is a tool used in the verification of design hypotheses as well as being a source of new ideas, new designs, and new hypotheses, it forms an invaluable link in the process of systems engineering.
References


Abstract:

The Internet has evolved in its current status starting from end to end principle as a backbone of design methodology. The end to end principle makes the core network simple and all the intelligence can be moved to network end points or hosts, making core network simplified and optimized for forwarding only. Content Delivery Network (CDN) can be regarded as an upcoming challenge to design a tightly integrated overlay which makes use of web caching, request routing, server load balancing and content aware services to optimize the content delivery. CDN is developed to overcome the performance problem of the Internet such as network congestion, server overload and low bandwidth and thus improves the service quality of Internet. A CDN system replicates/distributes content to a number of CDN nodes from the content provider (web servers) making an overlay network over the Internet. The content is provided to the user requesting for the web page from the closest CDN node. In this paper we first give the overview of CDN followed by a general architecture, a brief survey and comparison of few existing CDNs.

Keywords: Content Delivery Network, Replica server, CDN node, Distributed Nodes

1. Introduction

Content Delivery Networks (CDN) is an upcoming technology used to overcome the inherent limitations of Internet when accessing the web content. This technique improves the speed of accessing web content, maximizes bandwidth, reduces the load on the original server and maintains correctness through content replication and hence improves the internet performance. The word CDN was coined in 1998 and still a major thrust area of research in computer network. Research on the blistering areas of CDN like replica placement, request routing, resource optimization, content pricing, traffic congestion etc. are still in progress and it is believed that the successful solution to these problem will make the future CDN more popular. Most of the popular web sites in the internet use CDN system for the fast delivery of their embedded content. A number of commercial and academic CDN exists; some of them are discussed in this paper.

The rest of the paper is organized as follows. Section 2 describes the overview of CDN. In section 3 a general architecture of CDN is discussed. A brief survey of some of the existing CDNs is discussed in section 4. In section 5 we have compared some of the existing CDNs. Section 6 discusses a study on the performance of CDN. Finally in Section 7 we conclude our paper.

2. Overview of Content Delivery Network

A CDN is a collection of network elements arranged in internet for more effective delivery of data to end users. CDN uses the replica of the servers and places the replicas close to the clients to reduce latency and bandwidth consumption. The replica server store very selective set of content. CDN gives a lot of benefits to
the content providers. The content providers may be the web sites. The popular web sites use the CDN providers for the fast delivery of any digital content. The contents are replicated in the server according the user demand or it can be replicated in advance. The contents are also distributed among the various CDN nodes.replica servers created in the internet. The CDN provider delivers the content to the clients from the nodes in close proximity. This is done to maximize the hit ratio to 100%. CDN system does a number of tasks. It redirects the requests to the closest node to bypass congestion and delivers the content. It is responsible for content outsourcing and distribution services to replicate or cache content to distributed CDN nodes on behalf of its customer (content provider). It is also responsible for content negotiation services to meet specific needs of each individual user. It also manages the network components, handle accounting and monitor and report the content usage.

3. Architecture of CDN

![Figure 1 General architecture of CDN](image)

The general architecture of a CDN system can have a number of components. Figure 1 shows the general architecture of CDN system. Three basic components are the end users, the content provider and the CDN provider. Internal 3 components (Request Routing System, Distributed System and Accounting System) of the CDN provider are also shown in the figure.

The End user requests for a web page. The content provider is the customer of the CDN provider that has the contents to be delivered to the end user. Here we have taken an example of xyz.com. The CDN provider is a proprietary organization or company that provides infrastructure facilities to content providers in order to deliver contents in a timely and reliable manner. Content provider updates the content in the CDN server. CDN system creates an overlay network over the internet. The system consists of a number of CDN nodes that are either the replica server or the nodes where the content is distributed. Replica server is
the server which holds the replica of the resources but it may act as an authoritative reference for client response. The request routing system of the CDN provider routes the request to the closest CDN node. The distribution system moves content to replica servers. This system also communicates with the request routing system through the feedback to assist in the replica server selection process for client requests. The accounting system aggregates and distills the accounting information into statistics and content detail records for use by the CDN server and billing organization. The figure also shows a typical communication between these components. (1) The end user request for the content of www.xyz.com. The request goes to the xyz.com server. (2a) xyz.com sends the index page to the end user and (2b) the xyz.com server simultaneously sends the request to the content provider to provide the contents to the end user on its behalf. (3) The Content provider through the request routing system routes the request to a closest CDN node and that CDN nodes sends the embedded objects of the requested page.

4. Survey of existence CDNs

There are many commercial and educational CDN. Some of the popular existing CDNs are discussed in this section.

4.1 Akamai

This is one of the most successful and commercial CDN that hosts most of the popular websites of yahoo, monster, IBM etc. It also built its own DNS network for fast delivery of requested content by resolving the host name of the URL to IP address. This is developed by MIT. It is the market leader in content delivery. It owns more than 18000 servers over 1000 network in 70 countries.

4.2 SyncCast

SyncCast offers complete solutions from application development, Web hosting and Internet connectivity to deployment and systems integration. It provides solutions for delivering digital content and related data via the Internet and other media. SyncCast uses load balancing equipment for load balancing client traffic. SyncCast is also a partner with large companies like Microsoft and Dell. SyncCast’s clients include the Motion Picture Association of America, Walmart Music, Lions Gate Films, Microsoft, EMI Music Group, Technicolor and Billboard Radio.

4.3 Globix

Globix has more than 1200 customers. Globix provide four types of services: Network Services, Hosting Services, Managed Services, and Media Services. Globix services are flexible, scalable, and cost-effective. Globix also provides Media services to capture, store, host and distribute media content from live event production, encoding, presentation tools, and traffic analysis. Globix IP backbone connects the customers to the Internet via a high-capacity network, fully owned and operated by Globix.

4.4 Accellion

Headquarter of this privately held company is California. It provides large scale file delivery service. It is a distributed file storage and transmission infrastructure for enterprise applications. Accellion also provides online desktop and server backup and recovery solutions. Accellion customers are industries such as advertising/Media production, manufacturing, healthcare, consumer goods, higher education etc.

4.4 Coral
Coral is a free, peer-to-peer academic content delivery network. When a user wants to use Coral, the content publisher has to append “.nyud.net:8090” to the hostname in a URL. Clients are redirected to the nearby Coral Web caches. Coral Web caches cooperate to transfer data from nearby peers whenever possible, minimizing both the load on the origin Web server and the latency perceived by the user.

4.5 Codeen

Codeen is also an academic CDN developed at Princeton University, USA. It provides caching of Web content and redirection of HTTP requests. In Codeen users set their internet caches to a nearby high bandwidth proxy that participates in the Codeen system. Then the requests to that proxy are sent to the system where the file is cached. The file is forwarded to the proxy and then to the client.

5. Comparison between the existing CDNs

In this section we have compared some of the existing CDNs on the basis of their type, organization, request routing technique, caching technique, content outsourcing and content type. Table 1(a) and 1(b) shows the comparison. CDN can be organized in two different ways: Overlay approach and Network approach. In the overlay approach, CDN nodes in the network handle the distribution of specific content type. In the network approach, networking devices like routers and switches are equipped with the code for identifying the specific application type and for forwarding the request based on predefined policies. Request routing techniques are of six types: Global Server Load Balancing, DNS-based request routing, HTTP redirection, URL rewriting, Any-casting and CDN peering discussed in [2]. Caching techniques are of two types: Intra-cluster and inter-cluster caching. This is also defined in [2]. Content outsourcing can be performed in two ways: Cooperative pull based and non-cooperative pull based approaches. In cooperative pull based approach, the end user requests are directed to the closest CDN node though DNS redirection. In non-cooperative pull based approach end user requests are directed either through DNS redirection or URL rewriting technique.

<table>
<thead>
<tr>
<th>CDN Name and Type</th>
<th>CDN Organization</th>
<th>Routing Technique</th>
<th>Caching Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akamai (Commercial)</td>
<td>Network and Overlay Approach</td>
<td>DNS-based request routing</td>
<td>Intra and Inter-cluster caching</td>
</tr>
<tr>
<td>SyncCast (Commercial)</td>
<td>Network Approach</td>
<td>Global Server Load Balancing</td>
<td>Intra-cluster caching</td>
</tr>
<tr>
<td>Globix (Commercial)</td>
<td>Network and Overlay Approach</td>
<td>Global Server Load Balancing</td>
<td>Intra-cluster caching</td>
</tr>
<tr>
<td>Accellion (Commercial)</td>
<td>Network Approach</td>
<td>HTTP redirection</td>
<td>Inter-cluster caching</td>
</tr>
<tr>
<td>Coral (Academic)</td>
<td>Overlay Approach</td>
<td>DNS-based request routing</td>
<td>Intra and Inter-cluster caching</td>
</tr>
<tr>
<td>Codeen (Academic)</td>
<td>Overlay Approach</td>
<td>HTTP redirection</td>
<td>Intra and Inter-cluster caching</td>
</tr>
</tbody>
</table>

Table 1a CDN comparison

<table>
<thead>
<tr>
<th>CDN Name</th>
<th>Content Outsourcing</th>
<th>Content Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akamai</td>
<td>Non-cooperative pull-based</td>
<td>Static Content, Streaming media</td>
</tr>
<tr>
<td></td>
<td>Non-cooperative pull-based</td>
<td>Streaming Media</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Globix</td>
<td>Non-cooperative pull-based</td>
<td>Internet infrastructure and network services</td>
</tr>
<tr>
<td>Accellion</td>
<td>Non-cooperative pull-based</td>
<td>Large File transfer Services</td>
</tr>
<tr>
<td>Coral</td>
<td>Cooperative pull-based</td>
<td>Static content</td>
</tr>
<tr>
<td>Codeen</td>
<td>Cooperative pull-based</td>
<td>Static content</td>
</tr>
</tbody>
</table>

Table 1b CDN comparison

6. Performance Study

We have examined the performance of the CDN in our study. Performance of CDN can be measured in many ways. One piece of work on the performance found from the simulation of a specific web server log file that there is a diminishing return to placing more replicas so as to minimize total content delivery cost, and that a surprisingly small number of replicas (less then 5) is enough to significantly improve optimized performance gain (82%).[3]. It is also observed that placing replicas among a small number of carefully chosen candidates (top client receiving most content and a few client with the highest out degree) allowed us to achieve the same performance gain as placing replicas among all clients with reduced computation cost. To improve performance perceived by client, replicas are usually chosen from a large number of locations where client’s demands are highly aggregated. The performance can also be measured on the basis of number of requests offloaded from origin server, their impact of client perceived latency and their ability to efficiently load balance request amongst a set of CDN servers. Krishnamurthy et al. [12] examined briefly how content distribution servers improve latency when compared to throughput from the origin servers. In more recent work, on the performance of peer to peer swarming CDN system, an intensive simulation of basic P2P swarming CDN system under different critical parameters setting, the simulation results shows the P2P swarming system can perform well under high bandwidth connection, uniform startup distribution and appropriate block size etc. and the P2P swarm content delivery system can scale well with nodes size and content size [14].

7. Conclusion

To overcome the performance problem of the internet such as network congestion, server overload, low bandwidth that arise when accessing popular content by many users, CDN is developed. Caching and replication technique has improved the performance of CDN and hence of Internet in many ways. In this paper, we have discussed the general architecture of CDN, few existing commercial and existing CDN followed by a performance study of the system. With the emerging new form of Internet content and services such as video on demand that requires high bandwidth, CDN is becoming more popular and their customer is growing rapidly. Research on the blistering areas of CDN like replica placement, request routing, resource optimization, content pricing, traffic congestion etc. are still in progress and it is believed that the successful solution to these problem will make the future CDN more popular.
References


Abstract - Have you ever wondered feeling the beautiful smell of Swiss Alps when you are sitting thousands of kilometers away in New York? This seems unreal but exploiting nature is what humans' are good at. Nature has given us many gifts including five senses. Till now Internet inarguably one of the strongest communication mediums at present, uses all senses Viz. Vision, Hearing and touching but two are missing viz. the power of smell and taste. In this paper, we will be venturing into how the smell will be transferred over the Internet and how to secure the system from unwanted smells/compounds. The main focus will be on standardizing a transfer process and controlling the smell transfer at end points.

No one would like to produce Phosgene, Mustard Gas etc at an endpoint that can kill a person or produce smell of burning rubber that can give user a bad experience. This paper will also discuss how to use an Odor firewall to stop unwanted smells/compound formation at an endpoint. All this seems unrealistic but within few years these things will penetrate into our life marking a new revolution.

Index – Intelligent Systems, Odor Transfer, Human-Computer Interaction

I. Introduction
Secure Aroma Transfer outlines the concept of transferring smell over Internet in a secure manner. Following points are being covered in this paper:

1) **Science behind Odor:** In this, we have discussed about the biological concept behind smell and its reception in nose. This includes all the biological and physical concepts and nature in which smell needs to be handled.

2) **Smell System and Working:** The working of whole system is discussed in this section. An ad-hoc client-system architect is taken as a reference for discussions later.

3) **Database:** We have discussed the architecture of database required for pattern storage and handling of packets for transfer of pattern over Internet.

4) **Packet Structures:** We have discussed about the structure of packets that are required for handling and transference of data over the Internet securely.

5) The need for securing the system and various applications of the system will be discussed in the last part.

II. Science behind Odor

A. The Biological Insight

From single cell organisms to Multi-cell organisms, everyone is formed from amino acids. Amino Acids are building blocks of every substance. Our hairs are made of proteins which in turn are created using amino acids. Odors are also formed of Amino Acids. These odors acts as a stimulant to the 100 million receptors cell existing in our nose. The stimulation depends on many parameters which are defined later in this paper.

B. Odor physiology

Olfaction depends upon the interaction between the odor stimulus and the olfactory epithelium. The olfactory membrane is a sensitive area, covering 4 to 6 square cm in each nostril. Beneath the membrane is a mucus layer. The nerve cells or peripheral receptor cells that primarily sense odors and fragrances are located in the epithelium. Cilia extend from the nerve cells into the mucus layer, which greatly increases the potential receptor area. The cilia are thought to contain the ultimate olfactory receptors, which are specialized protein molecules. Specific anosmia may result from the inability to synthesize the appropriate protein. The receptor cells transmit impulses to the olfactory bulb located at the base of the front brain. At the bulb, fibers from the nose contact with other nerves, which travel on to various parts of the brain.
For a substance to be detected as an odor by the receptor cells, several criteria must be met:

1) The substance must be volatile enough to permeate the air near the sensory area.
2) The substance must be at least slightly water-soluble to pass through the mucous layer and to the olfactory cells.
3) The substance must be lipid-soluble because olfactory cilia are composed primarily of lipid material; and finally.
4) A minimum number of odorous particles must be in contact with the receptors for a minimum length of time.

Thus, above written characteristics will be responsible for proper production of Odor on Producer end. The odor should be easily sniffed by the nostrils and should activate the olfactory neurons.

C. Primary classes of Olfactory Stimulants

Based on psychological tests, seven primary classes of olfactory stimulants have been found to preferentially excite separate olfactory cells. These classes are:

1) Ethereal
2) Camphoraceous
3) Musky
4) Floral
5) Minty
6) Pungent
7) Putrid

The nervous system integrates the responses from a number of cells to determine the identity of the primary odor stimulus being received. The intensity of the perceived odor class is related to the number of receptors bound and the degree of excitation of the olfactory cells.

A more flexible way of presenting the primary odors to clarify the idea of complex odors is through the use of Henning's odor prism

III. Smell System – Insight in Bioinformatics

Making computers understand smell is just like making them understand a pattern. Various techniques can be employed to make a system understand smell. But for this we need to attach a smelling device or a sniffer to the computer. The complete smelling system (including the hardware and the software) proposed here is depicted by the figure below.

The various parts of the system are as follows:

1) Smell Receiver or Sniffer
2) Smell Producer
3) Smell Database
4) Smell Encoder/Decoder

All most all parts are present on both the receiver and senders’ systems. Both of them may be located miles apart.

1) Smell Receiver or Sniffer - On the sender side, a sensing system is required, which just like any other chemical composition detector will try to judge the chemical combination of the vapors. Though developing such a system can be very complicated, but there are no barriers in the world of science. This thing can be very similar to a smoke detector.

A very major success has been achieved by the scientists at NASA. They have developed what they call an Electronic Nose or the ENose. It's a device that can learn to recognize almost any compound or combination of compounds. It can even be trained to distinguish between Pepsi and Coke. Like a human nose, the ENose is amazingly versatile, yet it's much more sensitive. "ENose can detect an electronic change of 1 part per million." So with NASA going the ENose way our Smell receiver already seems to be on the line of production soon. The device connects to a PC, which then determines what smell the electronic nose has captured.

2) Smell Producer - A system will be attached to a device which is just like an atomizer which
using a few set of chemicals will be able to produce vapors. Now the smell which as discussed earlier is the composition of chemicals can be produced by a few reactions. Another, solution for the systems which intend to generate only a few types of smell, is to store the smells using particular chemicals. For example, a company having some ten types of perfume can always put very small samples in a machine at its store.

3) Smell Database - Now, the computer system in the model contains a database as backend. This smell database will store various chemicals of particular combinations which can be used to compose different type of smell. Even the database can hold record containing data regarding chemicals that can be used as whole or as substitute constituents to produce a required type of smell (if only the particular chemical are unavailable at the Smell Producer).

4) Smell Encoder/Decoder - It is responsible for the processing of input from the Smell Receptor. It will look in to the database to find suitable chemicals constituting a particular smell. And on the client side it will recommend the Smell Producer as which chemical combination to be used so as to produce the desired smell.

**Working**

Now as illustrated by the model, the user makes the request for scanning a smell. The required smell is sensed by the Smell Receptor and then the sensed signal is sent to the Encoder. The encoder will search for the exact smell first in the database, and if found then the chemical composition is extracted. If the exact smell chemicals are not found then it can suggests some options to the user. Now this part is pure A.I. and will require the learning capability in the system. Once the smell is recognized and the chemicals components have been sorted out, data will be sent to the client side. Once again here the decoder will look in its database as how can the required smell be produced. If an exact match is found, then the data is sent to the Smell Producer. If an exact match is not found, then based on intelligence the system will try to guess alternative chemical compositions. The middleware on the both ends will be responsible for letting the editing process in the database. The model here leaves all the calculation part on the encoder so as to make it work with a wide variety of interfaces thus reducing the complexity at the front end.

**IV. Database**

The odor structures and patterns will be stored in a database. The database will be stored at two locations viz. on the hardware and at a centralized server. The database stored at the centralized server will be updated from the information gained at individual hardware. Following is the detailed database structure that will be used.

**Table 1 – Pattern**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Index</td>
<td>Pattern Index will refer to the identity of the pattern</td>
</tr>
<tr>
<td>Ethereal Index</td>
<td>This attribute will refer to the index of the ethereal compound</td>
</tr>
<tr>
<td>Ethereal Intensity</td>
<td>This attribute will let know the device how intense is the concentration of ethereal compound.</td>
</tr>
<tr>
<td>Camphor Index</td>
<td>This attribute will refer to the index of the camphor compound</td>
</tr>
<tr>
<td>Camphor Intensity</td>
<td>This attribute will let know the device how intense is the concentration of camphor compound.</td>
</tr>
<tr>
<td>Musky Index</td>
<td>This attribute will refer to the index of the musky compound</td>
</tr>
<tr>
<td>Musky Intensity</td>
<td>This attribute will let know the device how intense is the concentration of musky compound.</td>
</tr>
</tbody>
</table>
Putrid Index: This attribute will refer to the index of the putrid compound.
Putrid Intensity: This attribute will let know the device how intense is the concentration of putrid compound.

Pungent Index: This attribute will refer to the index of the pungent compound.
Pungent Intensity: This attribute will let know the device how intense is the concentration of pungent compound.

Floral Index: This attribute will refer to the index of the floral compound.
Floral Intensity: This attribute will let know the device how intense is the concentration of floral compound.

Minty Index: This attribute will refer to the index of the minty compound.
Minty Intensity: This attribute will let know the device how intense is the concentration of minty compound.

Danger Level: The danger level is associated with each pattern. This danger level is important to disallow usage of certain compounds that are harmful.

To secure the odor transfer system, we need to perform some base security checks, one of which is to block certain IP’s that are repeatedly sending unwanted odor information. This unwanted odor information maybe unacceptable to the system and scenario in which it will be used or may be categorized as dangerous odor.

Following are the attributes associated with blacklisted IP table:

Index: Reference number to blocked IP will be stored here in this attribute.
IP: The IP blocked will be stored in this attribute in the table.
ClientName: This attribute will contain the hostname of the client to be blocked and will also contain some OS fingerprinting data so that any IP can be unblocked in case the IP assignment was dynamic.
Reason to block: The reason to block will be mentioned automatically. This may be repeatedly sending unwanted data or dangerous smell or something else.

<table>
<thead>
<tr>
<th>Blacklisted IPs Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
</tr>
<tr>
<td>IP</td>
</tr>
<tr>
<td>ClientName</td>
</tr>
<tr>
<td>Reason to Block</td>
</tr>
</tbody>
</table>

Fig 4: Table 2 – Blacklisted IPs

**Composition Tables**

Composition tables will contain the list of various chemical compounds specific to seven odor classes. These chemical compounds will later mix together to form various odors.

**Table 3 – Composition table for Odor classes**

CompoundID: The CompoundID field will contain a unique reference id specific to various chemical compounds listed in the table.

CompoundName: The CompoundName field will contain the name of the chemical compound referenced above.
Composition: The composition field will tell about the chemical composition of the compound. For example – If compound involved is methane, then simple CH₄ will be written in this tuple against composition field.

<table>
<thead>
<tr>
<th>ETHREAL Property Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>CompoundID</td>
</tr>
<tr>
<td>CompoundName</td>
</tr>
<tr>
<td>Composition</td>
</tr>
</tbody>
</table>

Fig 5: Table 3 – Ethereal Composition Tables – Similar tables need to be created for other odor classes

V. Packet Structure

Odor Transfer Protocol (ODTP) will be the protocol used to transfer data on the Internet. The protocol will support certain signals and packets that will provide status of the system and help in communication between two ends.

Following will be the packets that will be used to communicate over the network

- Login Packet
- Login OK Packet
- Odor Transfer Start Packet
- Odor Pattern Packet
- Odor Transfer Stop Packet
- Odor Transfer Status Packet

Data Packet:

<table>
<thead>
<tr>
<th>0</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODT P</td>
<td>Version Info</td>
</tr>
<tr>
<td>Packet Length</td>
<td>Service</td>
</tr>
<tr>
<td>Session ID</td>
<td>Status</td>
</tr>
</tbody>
</table>

Data (0 - 224) (Ethereal, Musky, Pungent, Minty, Putrid, Floral, Camphoraceous)

Fig 6: Data Packet

(ODTP Packet)

Above shown is the Data Packet that will contain the details of the smell to be sent. The data packet will consist of two parts viz. header and data. Header will be of 16 Bytes and will be common to each packet.

Header Details:

First four Bytes will point to the type of Protocol that will be used (in this case ODTP). Next four bytes will be used to display the protocol version. Next two bytes will be used to display the length of the payload. Next two bytes will be used to display type of payload viz. Initialization, Login, Odor Payload etc. Next three bytes will be used to display SessionId and last byte will be used to display connection status.

Data:

The Data field will be dynamic and the length will range from 0 to 28 Bytes. It will contain the code for odor that needs to produce at other end.

Login Packet:

This packet is used to make session between two machines by sending ClientID, ip, listening port and other details. Following listed are all the values that will be sent on both ends.

<table>
<thead>
<tr>
<th>Bytes</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ClientID</td>
</tr>
<tr>
<td>4</td>
<td>IP</td>
</tr>
<tr>
<td>2</td>
<td>Listening Port</td>
</tr>
<tr>
<td>1</td>
<td>Encryption</td>
</tr>
<tr>
<td>1</td>
<td>Compression</td>
</tr>
<tr>
<td>2</td>
<td>Max Transfer Speed</td>
</tr>
<tr>
<td>2</td>
<td>Min Transfer Speed</td>
</tr>
<tr>
<td>2</td>
<td>Min Time Interval</td>
</tr>
</tbody>
</table>

Login OK Packet:

This Packet will be send by peer on other side in reply to the Login packet. It will contain information about other end’s IP, listening port for further communication, ClientID, support for encryption and compression etc.

<table>
<thead>
<tr>
<th>Bytes</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ClientID</td>
</tr>
<tr>
<td>4</td>
<td>IP</td>
</tr>
<tr>
<td>2</td>
<td>Listening Port</td>
</tr>
<tr>
<td>1</td>
<td>Encryption</td>
</tr>
<tr>
<td>1</td>
<td>Compression</td>
</tr>
<tr>
<td>2</td>
<td>Max Transfer Speed</td>
</tr>
</tbody>
</table>
**Odor Pattern Packet:**

<table>
<thead>
<tr>
<th>PatternID</th>
<th>ObjectName</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Class</td>
<td>Composition</td>
</tr>
<tr>
<td>Intensity(in %)</td>
<td></td>
</tr>
<tr>
<td>Primary Class</td>
<td>Composition</td>
</tr>
<tr>
<td>Intensity(in %)</td>
<td></td>
</tr>
</tbody>
</table>

Fig 7: Odor Pattern Packet

The Odor pattern packet will contain the information about the percentage intensity, composition and odor class type. The final smell to be produced will be labeled with a name of the substance it smells like.

**Odor Transfer Start Packet:**

This packet is used to start transfer of odor from other end. This packet will contain only header with payload length zero and service value with code for start odor.

**Odor Transfer Stop Packet:**

This packet is used to stop transfer of odor from other end. This packet will contain only header with payload length zero and service value with code for stop odor.

**Transfer Status Packet**

This Packet will be used to send status of a client to peer on other side. This packet will contain information about odor receive, odor send, how many new smells it got and send in last hour.

**VI. Securing the Smell System**

The smell system in itself is a very unique idea that will change the way we do things. But securing the system will be an utmost important thing to do. If some malicious being is successful in transferring Phosgene over the Internet, he can kill so many peoples. To minimize the danger the concept of Odor Firewalls will come into being. Moreover, many other considerations as discussed below should be taken into consideration.

a) **BlackListed IP’s** – A BlackListed IP table is used to maintain the information of blocked IP’s in database. This table will contain ClientName (which will be unique for each machine [hardware] and will not change even in case of dynamic IP’s), IP and port information. The machines from which something malicious, that may be transferring dangerous odor packets periodically, is being done will be blocked.

b) **Payload Integrity Check** – An integrity check will be done on all the packets that need to be transferred. The techniques that can be used may be hashing or CRC check.

c) **Dangerous Odor Check** – All the odors to be transferred will be first checked for Danger Level of the smell. The danger level if large will not allow encoding of the smell on the sender end.

d) **Database Security** – The database structure can only be modified/update by a service running on a web server that will be responsible for updation of database on clients.

e) **Hardware Intelligence in Atomizer** – At hardware level there is a consideration on Physical Environment Dependency that needs to be taken care of: Whenever odor production will start on the producer end, the hardware should check the outside environment chemical composition and should check if the odor to be produced by the atomizer can react with the environment and result in drastic after-effects.

f) **Hardware Agreements** – All the interaction between two ends shall depend on the negotiations performed before starting the transfer. The negotiation will involve encryption, compression information.

g) **Devices of different Brand** – In case, the two interacting entities have different brand of devices, there may occur a situation where one entity encode a smell but the other end decode it as something different. In that regard, a standardization of the hardware needs to be done.

**VII. Applications**

1) Consider the case of a perfume company. All it needs is to set an outlet, which provides intelligent solutions, which will require only the above discussed Smell Model. The customer will ask for the smell of a particular perfume from the system by accessing the front end. And the system will produce the smell on the spot. Then
depending upon the decision of the customer the product can be ordered at the online store.

2) This technology can be used in mission critical sensitive areas such as sophisticated research facilities. For examples, if the amount of ammonia grows in the air, the system will automatically activate an alarm. This is an idea similar to the detection of amount of ammonia in space crafts, the main aim behind Enose by NASA.

3) Consider an online food order system. User just needs to visit the portal and click on his favorite coffee. He can smell the coffee and can order the coffee online. This can be done with all the eatables and once it is done, online shopping will change forever.

4) Similarly as till now feelings have been transferred over the internet in the form of text, sounds and pictures, smell can also come into picture. Imagine a greeting card which not only has pleasing audio-visuals but also produces a soothing sweet smell. Even user defined smells could be transported. A girl could then make her boyfriend feel special by sending smell laden greeting card.

5) This system can bring new realism to games; imagine smelling blood in a bloody game like Resident Evil or the smell of leaves when you pass through a forest in a game like Jurassic Park.

6) Even for many other areas can benefit from smell over internet. Invent a new smell and then transport it through the internet.

7) Chatting will be the biggest opportunity domain for Smell Systems. The smell system will produce the smell of one end on the other end thus virtually bringing people closer.

VIII. Conclusion

Think of surfing Domino’s website and clicking on your favorite pizza for order. With a single click, one can smell a very pleasant smell of the pizza in the room and can change his order or finalizes his order. Integrating smell with computers is not a far fetched impossible concept. Smell promises to add a new dimension to the internet. And in a matter of few years this technology will find place in normal Personal Computers similar to audio-video systems. The power of smell really rocks. And from now multi-media will contain one more media, the ‘Media of Smell’.

IX. References

[1] Science of Odor - A
http://www.extension.iastate.edu/Publications/PM1963A.pdf

http://www.extension.iastate.edu/Publications/PM1963B.pdf

http://www.extension.iastate.edu/Publications/PM1963C.pdf

http://www.extension.iastate.edu/Publications/PM1963D.pdf


[6] ENose
http://science.nasa.gov/headlines/y2004/06oct_enose.htm
Abstract

The Resource Description Framework is a W3C standard language for representing information about resources in the world wide web; RDF provides a common framework for expressing this information, so it can be exchanged between applications without loss of meaning. In this paper I used RDF and RDFSchemas to represent Dravidian University information database to facilitate students for getting the required information. In this paper, I used SPARQL as a query language to retrieve information from database.

Keywords

Semantic, RDF, RDFSchema, SPARQL

Introduction

The semantic web vision was conceived by Tim Brener-Lee, the inventor of World Wide Web (WWW). Semantic web is an extension of the current web, in which information is given well-defined meaning. Semantic in the ‘semantic web’ is not that the computers can understand the meaning of anything, but the logical pieces of meaning can be mechanically manipulated by a machine to useful ends. In this paper for stressing of the semantic web, I used Resource Description Framework (RDF) for knowledge representation and SPARQL, a query language for information retrieval. For this purpose, I took a case study of the Dravidian University information. Starting with the description of RDF and RDFSchemas, I analyzed the representation of Dravidian University information and worked on sample queries by using SPARQL for retrieving the information. The paper aims to highlight the importance of semantic web by using the available tools like RDF, RDF Schema and SPARQL.

RDF and RDF Schema

In February 2004, The World Wide Web Consortium released the Resource Description Framework (RDF) as W3C Recommendation. RDF is used to represent information and to exchange knowledge in the Web. It gives you a way to make information machine-processable. RDF uses a general method to decompose knowledge into pieces called triples. The triples can be represented as subject, predicate and object. In RDF, the English statement
“Tim Berners-Lee invented World Wide Web”

Could be represented by RDF statement having

- A subject *Tim Berners-Lee*
- A predicate *invented*
- And an object *World Wide Web*

RDF statements may be encoded using various serialization syntaxes. The RDF statement above would be represented by the graph model as shown below

![RDF Graph]

Subject and objects are represented by nodes and predicate is represented by an arc. RDF’s vocabulary description language, RDF Schema, is a semantic extension of RDF. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDF Schema vocabulary descriptions are written in RDF using the terms described in this document. These resources are used to determine characteristics of other resources, such as the domains and ranges of properties.

**Representing Facts in RDF and RDFS**

The information is categorized under four resources such as department, courses in each department, employees in the departments and research activities of each employee. The RDF Schema of our domain uses simple namespaces like

- Dept – represents departments in the university
- Course - courses in each department
- Staff - employees in each department
- Projects - research activities of each employee in the university

The ‘Dept’ class represents all the departments in the university. For ex: “computer science head C.lokanatha reddy”. In this statement computer science is a subject, head is a predicate and ‘C.Lokanatha Reddy’ is Object. Like this I represented all the departments with concerned heads and year of establishment they started. Here ‘depid’ is an identification code for each department. Similarly the Course class represents all courses that are being offered department-wise. For ex: if the computer science department is offering M.C.A course then this fact can be represented as “depid offering M.C.A” (depid -> subject offering -> predicate M.C.A->object). Similarly the Course resource represents other details such as eligibility to join, admission process whether through entrance test or interview by the university, pattern of the course i.e., yearly or semester, fee particulars and category of the course i.e., degree or PG or Research etc.. After understanding the meaning of RDF and RDF Schema, let us now
proceed to see how these two work, by taking the Dravidian University information as a case study.

The ‘Staff’ class represents the information about employees working in each department. This class has a relation with the ‘Dept’ with depid. This class contains the details of employee name, qualification, experience and address. Finally, the ‘Project’ class is related to the research activities in the university. This has connected with Staff class through ‘empid’. This contains information about research projects of each employee in the department and details of funding agency. The detailed graph is represented as above. Circles and rectangles represent subject and object and arcs represent predicate.

**Querying with SPARQL**

SPARQL is a query language for getting information from the RDF Graphs. There are several tools and Application Programming Interface (APIs) that already provide SPARQL
functionality. I made an attempt with ARQ – SPARQL processor for jena for information retrieval.

The example queries are listed below

1) List the departments in the university

```
prefix ex: <http://dravidianuniversity.org/>  
prefix xsd: <http://www.w3.org/2001/XMLSchema#>  
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  

SELECT ?name ?head ?established  
WHERE {  
  ?n ex:depid ?id;  
  ex:depname ?name;  
  ex:depestablished ?established;  
  ex:dephead ?head.  
}  
ORDER BY ?name
```

2) List the Courses offered by “Computer Science” Department

```
prefix ex: <http://dravidianuniversity.org/>  
prefix xsd: <http://www.w3.org/2001/XMLSchema#>  
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  

WHERE {  
  ?n ex:depid ?id;  
  ex:depname "Computer Science".  
  ?m ex:depid ?id;  
  ex:coursename ?cname;  
  ex:eligibility ?eligibility;  
  ex:admission ?adm;  
  ex:pattern ?pattern;  
  ex:fee ?fee. }
```

3) List the Employees in each department

```
prefix ex: <http://dravidianuniversity.org/>  
prefix xsd: <http://www.w3.org/2001/XMLSchema#>  
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  

WHERE {  
  ?n ex:depid ?id.  
  ?m ex:depid ?id;  
  ex:empname ?name;
```
4) List the PG courses offered in each department

```sql
prefix ex: <http://dravidianuniversity.org/>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

WHERE {
  ?n ex:depid ?id;
  ex:deptname ?name
  ?m ex:depid ?id;
  ex:coursename ?cname;
  ex:eligibility ?eligibility;
  ex:admission ?adm;
  ex:pattern ?pattern;
  ex:category “PG”;
  ex:fee ?fee. }
```

5) List the projects done by employees in each department

```sql
prefix ex: <http://dravidianuniversity.org/>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

WHERE {
  ?n ex:depid ?id;
  ex:deptname ?dname
  ?m ex:depid ?id;
  ex:empname ?ename;
  ex:projectname ?prname;
  ex:funding ?funding .}
```

**CONCLUSION AND FURTHER WORK**

In this paper an attempt is made to discuss how the information can be represented using RDF and RDF Schemas and how it can be retrieved it using SPARQL. This is entirely different from existing searching tool, as the existing searching engines are capable of retrieving information with keywords. That may result in many inappropriate results. But by using RDF and SPARQL inappropriate results can be minimized to a great extent. Though the appropriateness of the usability of the tool is significant, it has practical difficult in developing the code. Unlike the other search engines the present tool needs two output
formats, one for designing and another for meaning of the content. However, the merit of usability over shadows the difficulty with regard to code. The usability can be extended by adding additional information about the university and making front-end to facilitate query data for user.

References


[4] where to get SPARQL : C:\ARQ\doc\download.html


Cooperative Intrusion Detection for Detecting Novel Attacks Using Real Time Data Mining Approach

R.Pushparaj, Final Year M.E (CSE), S.Bhuvana M.E, Lecturer, CSE. VLB Janakiammal College of Engineering and Technology, Coimbatore, TamilNadu, India. 

pushparaj_vlb@yahoo.co.in , pusheager@sify.com

Abstract

The Internet is growing everyday and the growth appears to be accelerating. As more numbers of users become connected to Internet provides window of opportunity for the malicious users to do their damage becomes very high and lucrative. We proposed a Cooperative Intrusion Detection System (CIDS) approach that combines the advantages of low false-positive rate of signature-based intrusion detection system (IDS) and the ability of anomaly detection system (ADS) to detect novel unknown attacks. By mining anomalous Internet traffic episodes from Internet connections, we are going to build an ADS that detects anomalies beyond the capabilities of signature-based SNORT or Bro systems. A weighted signature generation scheme is developed to integrate ADS with SNORT by extracting signatures from anomalies detected. CIDS extracts signatures from the output of ADS and adds them into the SNORT signature database for fast and accurate intrusion detection. We hope that the signatures generated by ADS will upgrade the SNORT performance by some percent. The CIDS approach proves the vitality of detecting intrusions and anomalies, simultaneously, by automated data mining and signature generation over Internet connection episodes.

Keywords: Network security, intrusion detection systems, Internet traffic data mining, frequent episode rules, anomaly detection, anomaly and normality scores, signature generation, SNORT systems.

1. INTRODUCTION

Security of network systems is becoming increasingly important as more and more sensitive information is being stored and manipulated online. Intrusion Detection Systems (IDSs) have thus become a critical technology to help protect these systems. Intrusions and anomalies are two different kinds of attacks in an open network environment. An intrusion takes place when an unauthorized access of a host computer system is attempted. An anomaly is observed at the network connection level. Both attack types may compromise valuable hosts, disclose sensitive data, deny services to legitimate users, and pull down network-based computing resources. The intrusion detection system (IDS) offers intelligent protection of networked computers or distributed resources much
better than using fixed-rule firewalls. To protect *telnet*, *http*, *ftp*, *SMTP*, *pop3*, *Email*, and *authentication* services, early detection of Internet anomalies in routers, gateways, hosts, and servers are a necessity.

1.1 Intrusion versus Anomaly Detection Systems

In this paper, we consider signature-matching *intrusion detection systems* (IDS), based on the *misuse model*. We design *anomaly detection system* (ADS) based on the *anomaly-based model* characterized in Table 1. The misuse IDS model matches attack signature with pre-stored signatures from known attacks. The signatures are manually constructed by security experts analyzing previous attacks. The collected signatures are used to match with incoming traffic to detect intrusions. These are conventional systems that detect known attacks with low false alarms. The major problem with this approach is that these IDSs fail to generalize to detect new attacks or attacks without known signatures. Furthermore, signature matching performs well only for single-connection attacks. With the sophistication of attackers, more attacks involve multiple connections. This limits the detection range by signature matching.

On the other hand, an anomaly-based system uses a different philosophy. It treats any network connection violating the normal profile as an anomaly. A network anomaly is revealed if the incoming traffic pattern deviates from the normal profiles significantly. Through a data mining approach, anomaly detection discovers temporal characteristics of network traffic. This system can detect unknown attacks and handles multiconnection attacks well. However, anomaly detection may result in higher false alarms.

### Table 1. Comparison of Misuse Model and Anomaly Model for Intrusion Detection

<table>
<thead>
<tr>
<th>Attack Characteristics</th>
<th>Misuse Intrusion Detection System (IDS)</th>
<th>Anomaly Detection System (ADS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profiling level</td>
<td>Packet scanning to reveal misuse, abnormal, and guilty patterns</td>
<td>Mining TCP/UDP/ICMP connections to save normal traffic profiles</td>
</tr>
<tr>
<td>Detection mechanisms</td>
<td>Signature matching with known attack patterns and misuse logs</td>
<td>Profile matching with association or episode rules against normal traffic</td>
</tr>
<tr>
<td>Implementation requirements</td>
<td>Signature database and signature update mechanisms</td>
<td>Mining of audit data, episode rule generation, and database update</td>
</tr>
<tr>
<td>advantages and shortcomings</td>
<td>Detecting known attacks with low false alarms, fail to detect unknown attacks</td>
<td>Detect all attacks with higher false alarms, adaptive to network changes</td>
</tr>
<tr>
<td>Representative Systems</td>
<td>Snort [35], JAM [22], ADCPAR [13], MADAM ID [23]</td>
<td>Fuzzy Logic IDS [4], ADAM [3], EMERALD [31]</td>
</tr>
</tbody>
</table>

Existing IDSs are built with either signature-based or anomaly-based systems. The design philosophies of these two models are quite different, and they were rarely mixed up in existing IDS products from the security industry.
In this paper, we present a new co-operative intrusion detection system (CIDS) which is network based. This system combines the positive features of both intrusion detection models to achieve higher detection accuracy, lower false alarms, and, thus, a raised level of cyber trust. An adaptive base support threshold is applied on selected axis attributes in mining the Internet episode rules. The episode rules are used to build the CIDS, which detects not only known intrusive attacks but also anomalous connection sequences.

The rest of the paper is organized as follows: Section 2 reviews related works and distinguishes the new approach from previous solutions. Section 3 introduces traffic data mining and describes the CIDS architecture. We present Internet episode rules and pruning techniques in Section 4. The weighted signature generation is specified in Section 5. Finally, in Section 6, we summarize the contributions and comment on further research needed.

2 RELATED WORKS AND OUR APPROACH

In the past, data mining techniques such as using association rules were suggested to build IDS [1], [5]. Lazarevic et al have distinguished the differences between single-connection and multiconnection attacks. Both signature-based and anomaly-based IDSs are sensitive to the attack characteristics, system training history, services provided, and underlying network conditions. Data mining techniques are also used to build classification models from labeled attacks [8], [9]. SNORT [4] and Bro [14] are two widely used IDSs that are based on the misuse model. Intrusion detection must be designed to monitor the connection features at the network, transport, and application layers [3], [6]. The MIT/LL IDS evaluation data set and reported IDS performance results were analyzed in [10], [11]. We use this attack data set with mixed background traffic to test the effectiveness of CIDS. The concept of frequent episode rules (FERs) was first proposed by Mannila and Toivonen [12]. Subsequently, Lee suggested a framework to specify FERs for anomaly detection against normal traffic profiles. They developed a level wise data mining algorithm for building ADS. Fan extended Lee’s work to discover accurate boundaries between known attacks and unknown anomalies. in and Hwang [13] refined the rule formulation procedure with an adaptive base-support algorithm to mine normal traffic records. Different axis attribute values apply different thresholds. Kaleton Internet [7] built a prototype system by combining the two detection systems, but they work independently without interactions. We consider close cooperation between the two subsystems.

In this paper, we propose the CIDS architecture. The CIDS integrates the flexibility of ADS with the accuracy of signature-based IDS. ADS is designed by mining FERs [8], [12] over Internet connections. We have developed a new weighted signature generation algorithm to characterize anomalous attacks and extract their signatures. The new signatures are generated from anomalies detected by ADS. This idea was inspired by earlier works on weighted association rules [15], [16]. This new approach automatically enables CIDS to detect similar anomalous attacks in the future.
3 Co-operative Intrusion Detection System (CIDS)

In this section, we first introduce the data mining concept for hybrid intrusion and anomaly detection. Then, we describe the CIDS architecture, the ADS design, and the connection features used in ADS and automated signature generation.

3.1 Traffic Data Mining for Network Anomaly Detection

Open networks face threats from both system intrusions and anomalies in Transmission Control Protocol (TCP), User Datagram Protocol (UDP), or ICMP connections.

Fig. 1. Data mining scheme for network anomaly detection over Internet connection records.

Fig. 1 shows the three major components of our network anomaly detection process. First, we apply the normal profile database and construct the anomaly detection engine. The detection engine is capable of detecting anomalous episodes that are caused by traffic anomalies. The connection records are extracted from audited Internet traffic. The concept of FER will be covered in Section 4. The episode rule mining engine consists of two phases of development. The training phase is needed to generate the normal traffic database without attacks. Attacks may appear in the detection phase. The anomaly is detected once the episode rule describing the real traffic connections cannot find any match with normal connection rules in the database. With this reasoning, the network anomaly is detected by a normal-use detection model. We have generated an attack data set by a mixture of locally captured Internet trace files and the DARPA 1999 IDS evaluation data set. We stretched the trace file to mix with the Massachusetts Institute of Technology/ Lincoln Laboratory traffic files collected. We use the toolkits by Mahoney and Chan to mix Internet traffic data with the MIT/LL data set. A drawback of the FER-
based approach is caused by the fact that many attacks are triggered by a single connection and may not generate anomalous FERs. In order to solve this problem, we keep a keen interest on rare attributes of single connections. For example, connections with the same source and destination addresses are often attacks. Another problem is that a single attack may last for a long period of time. To solve this problem, we use connection sequence numbers, instead of time stamps, to mine connections heading to the same destination.

3.2 The CIDS System Architecture

Anomaly-based systems are supposed to detect unknown attacks. These systems are often designed for offline analysis due to their expensive processing and memory overheads. Signature-based system leverages manually characterized attack signatures to detect known attacks in real-time traffic. The CIDS illustrated in Fig. 2 integrates the flexibility of ADS with the accuracy of a signature-based SNORT. The SNORT is connected in cascade with the custom-designed ADS. These two subsystems join hands to cover all traffic events initiated by both legitimate and malicious users. By 2004, SNORT has accumulated more than 2,400 attack signatures in its database.

Fig. 2. A co-operative intrusion detection system built with a SNORT and an anomaly detection subsystem (ADS) through automated signature generation from Internet episodes.
In CIDS operations, the first step is to filter out the known attack traffic by SNORT through signature matching with the database. The remaining traffic containing unknown or burst attacks is fed to the episode-mining engine to generate frequent episode rules with different levels of support threshold. This leveling allows the detection of some rare episodes, declared as anomalies. The frequent episodes are compared with precomputed frequent episodes from normal traffic. The episodes that do not match the normal profiles or match them with unusually high frequency are labeled as anomalous. The anomalous episodes are used to generate signatures which capture the anomalous behavior using a weighted frequent item set mining scheme. These signatures are then added to the SNORT database for future detection of similar attacks. Unknown, burst, or multiconnection attacks are detectable by ADS. The signature generation unit bridges two detection subsystems in the shaded boxes. This unit characterizes the detected anomalies and extracts their signatures. We built an ADS by using the FER mining mechanisms to be described in Section 4. The new CIDS detects many novel attacks hidden in common Internet services, such as telnet, http, ftp, SMTP, e-mail, authentication, and so forth. The CIDS deployment appeals particularly to protect network-based clusters of computers, resources inside internal networks (intranets), and computational Grids.

3.3 Internet Connection Features

The performance of ADS is directly affected by the features used in training and rule generation. Lee and Stolfo and Lazarevic used connection features, temporal statistics, and content features for building IDSs. Table 1 summarized the connection and temporal statistics features used in our ADS to generate FERs by training and testing.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature Name</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Connection Features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>timestamp</td>
<td>Time when the connection begins</td>
<td></td>
</tr>
<tr>
<td>duration</td>
<td>Duration of the connection in second</td>
<td></td>
</tr>
<tr>
<td>ip_proto</td>
<td>IP protocol type</td>
<td></td>
</tr>
<tr>
<td>src_ip</td>
<td>Source IP address</td>
<td></td>
</tr>
<tr>
<td>dst_ip</td>
<td>Destination IP address</td>
<td></td>
</tr>
<tr>
<td>service</td>
<td>Network service on the destination, e.g., http, ftp</td>
<td></td>
</tr>
<tr>
<td>icmp_type</td>
<td>ICMP message type</td>
<td></td>
</tr>
<tr>
<td>src_types</td>
<td>Bytes sent by the source</td>
<td></td>
</tr>
<tr>
<td>dst_types</td>
<td>Bytes sent by the destination</td>
<td></td>
</tr>
<tr>
<td>flags</td>
<td>SF: Both SYN and FIN packets are known, S0: Only SYN packet seen in a connection, REJ: Connection rejected by the destination</td>
<td></td>
</tr>
<tr>
<td><strong>Temporal statistical features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>src_count</td>
<td>Number of connections from the same source</td>
<td></td>
</tr>
<tr>
<td>dst_count</td>
<td>Number of connections to the same destination</td>
<td></td>
</tr>
<tr>
<td>service_count</td>
<td>Number of connections for the same service</td>
<td></td>
</tr>
<tr>
<td>syn_dst%</td>
<td>% of connections with same feature and SYN errors</td>
<td></td>
</tr>
<tr>
<td>service_dst%</td>
<td>% of connections per service to the same destination</td>
<td></td>
</tr>
<tr>
<td>syn_service%</td>
<td>% of connections with SYN error to the same port</td>
<td></td>
</tr>
<tr>
<td>ave_duration</td>
<td>Average duration of connections for the same service</td>
<td></td>
</tr>
<tr>
<td>ave_src_bytes</td>
<td>Average bytes sent by the source.</td>
<td></td>
</tr>
<tr>
<td>ave_dst_bytes</td>
<td>Average bytes sent by the destination.</td>
<td></td>
</tr>
</tbody>
</table>
Since we aim at detecting anomalies of network traffic, the content features, which are mainly extracted from system log files, are not used in this work. The connection features and temporal statistics will be used in CIDS construction. Connection level features are extracted from raw TCPdump files. They are used in both FER and signature generation. The flags are used to signal special connection status. We listed three different flags: SF, S0, and REJ.

Temporal statistics are related to connections with the same reference features. They can be used to improve the accuracy of signature generation. For example, by tracking the number of connections initiated from the same source, the Source_Count could be used to set the threshold when we generate new signatures.

4. INTERNET EPISODE RULES AND PRUNING TECHNIQUES

An Internet episode is represented by a sequence of connection events, such as TCP, UDP, ICMP, or other connections. An episode can be generated by legitimate users or malicious attackers. Frequent episodes mostly resulted from normal users. A rare episode is likely caused by intruders. Our purpose is to build an ADS that can distinguish the rare or abnormal episodes from the normal or frequent episodes automatically.

Fig. 3. Generation of a frequent episode rule by scanning stream of Internet traffic connection events from left to right using a small scanning window.

4.1 Generation of Internet Episode Rules

In Fig. 3, we show typical stream of Internet traffic, represented by a sequence of connection events labeled as E1, E2, E3, and so forth. These connection events are related to various Internet service commands such as http, ftp, SMTP, authentication, and so forth. Note that some events may repeat to appear in the sequence. The time instants of these connections, in seconds, are marked below the events. A frequent episode is a set of connection events exceeding the occurrence threshold in a scanning window. A FER is generated out of a collection of frequent episodes. The FER is defined over episode sequences corresponding to multiple connection events in a roll.

For the first window in Fig. 3, we cover a sequence of three connection events: E2, E1, and E3. The event E2 triggers the occurrence of events E1 and E2 in a cascade. This leads to the following FER on these three events. This rule is detected within a window size w.

The rule is backed by a support base s, confidence level c, and minimal occurrence f.
\( E_2 \rightarrow E_1, E_3 \) (s, c, w, f) \rightarrow (1)

s \rightarrow \text{refers to the probability of event } E_2 \text{ to occur,}

c \rightarrow \text{refers to the probability of the joint connection event } (E_1, E_3) \cup E_2 \text{ to take place after event } E_2.

We calculate,

\[
\begin{align*}
s &= \text{Prob} [E_2] \\
c &= \text{Prob} [(E_1, E_3) \cup E_2] / \text{Prob} [E_2]
\end{align*}
\]

The support value \( s \) reflects the percentage of minimum occurrences of the episode rule out of the total number of connection records audited. The confidence level \( c \) is the joint probability of the minimal occurrence of the joint episodes out of the support of the left-hand side (LHS) episode. The window size \( w \) is the scanning period of the window. The minimal occurrence \( f \) indicates the minimum number of occurrences to establish the rule in question.

For a real-life example, event \( E_2 \) could be an authentication service requested at time zero, presented by two attributes (service =) authentication; flag = SF), where the flag is defined in Table 1. Events \( E_1 \) and \( E_3 \) correspond to two consecutive SMTP service requests denoted by: (service =smtp) (service = smtp). We obtain the following FER with a confidence level \( c = 80 \) percent for an authentication service followed by two SMTP services detected within a scanning window \( w = 4 \) s. The three joint traffic events account a support level \( s = 10 \) percent out of all possible network connections. The minimal occurrence \( f = 1; 200 \) implies that 1,200 is the minimum number of occurrences for the rule to be generated into the rule base.

\((\text{service} = \text{authentication})\longrightarrow (\text{service} = \text{smtp}) (\text{service} = \text{smtp}) (0:8; 0:1; 4 s; 1; 200) \longrightarrow (2)\)

4.2 A Base-Support Data Mining Scheme

Most mining techniques exclude infrequent traffic patterns. This will make the IDS ineffective in detecting rare network events. If we lower the support threshold, then a large number of uninteresting patterns associated with frequent services will be discovered. We introduce a new base support mining process to handle this problem. The process is specified in Algorithm 1. Our method is improved from the level wise algorithm by Lee.

**Algorithm 1. Base-Support Traffic Data Mining Algorithm**

1: INPUT: Base-support threshold \( f_0 \), all axis attributes and the set \( T \) of all network connections
2: OUTPUT: New FERs to add into existing rule set \( L \)
3: for each axis item set \( X \) in \( T \), do
By using Lee’s algorithm, one iteratively lowers the minimum support value. Initially, a high minimum support value is chosen to find the episodes related to high frequency axis attribute values. Then, the procedure iteratively lowers the support threshold by half. This links each new candidate FER with at least one new axis value. The procedure terminates when a very small threshold is reached. Let $X$ be an item set. The support base of $X$ denoted by $\text{sup base}(X)$ is the support value of the axis item set. For example, when choosing the service and flag as the axis attributes, the support base for item set $X = (\text{service} = \text{ftp}; \text{flag} = S0; \text{srchost} = 128:1:1:1; \text{destination} = 121:1:1:1)$ is defined by $\text{sup base}(X) = \text{support} (\text{service} = \text{ftp}; \text{flag} = S0)$. The base-support fraction $f$ for item set $X$ is defined by:

$$F(X) = \text{support}(X) = \text{base sup}(X): \text{--------> (3)}$$

Similarly, the base-support fraction of an episode is defined as the percentage of the number of minimal episode occurrences to the total number of records in $T$, which contains the most uncommon axis attributes embedded in this episode. The minimum support base value of an episode $e_1; e_2; \ldots; e_n$ is denoted by $\text{min \{base sup (ei)\}}$.

### 4.3 Episode Rule Training from Normal Traffic

Fig. 4 shows the FER generation and rule-matching process in anomaly detection based on Algorithm 1. When attacks are detected by SNORT, their time stamps are passed to the packets eliminator, and the corresponding traffic flows are deleted. The rest of the traffic is passed to the ADS. When a FER generated from the traffic does not match the normal FER database, an unknown FER anomaly is suspected.

When the matched rule occurs beyond the threshold, multiple FER anomalies are suspected. The FER anomalies are confirmed by checking some error flags and temporal statistics listed in Table 1. Otherwise, the traffic connection is considered normal. To generate FERs for normal traffic profiles, the attack free training connection records are fed into the data mining engine. We use the audit data sets collected in weeks 1 and 3 of the 1999 MIT/LL IDS evaluation package.
Fig. 4. Matching with frequent episode rules to detect anomalies in incoming traffic.

We generated 92 FERs with the limited training time. We do not use FERs with extremely low support values. After finding FERs from each day’s audit record, we simply merge them into a large rule set by removing all redundant rules based on the pruning techniques discussed in Section 4.4.

4.4 Pruning of Ineffective Episode Rules

We consider a FER effective if it is applicable and more frequently used in the anomaly detection process. An episode rule is ineffective if it is rarely used in detecting anomalies. Some FERs differ only at the LHS or at the RHS. Keeping all rules generated will enlarge the search space and thus increase the overhead. The following FER transformation laws will reduce the rule search space significantly.

4.4.1 Elimination of Redundant Episode Rules

In general, rules with shorter LHSs are more effective than rules with longer LHSs. This is because shorter rules are often much easier to compare. For example, in the following rule,

\[(\text{Service} = \text{http})(\text{Service} = \text{authentication}) \rightarrow (\text{service} = \text{smtp}) \space (0:6; \space 0:1) :\rightarrow \space (4)\]

The rule above is considered ineffective with the existence of the following rule:

\[(\text{service} = \text{authentication}) \rightarrow (\text{service} = \text{smtp}) \space (0:65; \space 0:1) :\rightarrow \space (5)\]
The authentication is related only to the smtp operation; the http does not affect the other two item sets. Therefore, (service = http) can be ignored. Longer rules may introduce some redundant information. Removing them from the normal profile will reduce the false alarms.

4.4.2 Reconstruction of Episode Rules

Many FERs detected from the network traffic have some transitive patterns. Suppose we have two rules A → B and B → C in the rule set. Then, the longer rule A → B, C is implied. Since we reconstruct this rule from two shorter rules, the longer rule A → B, C becomes redundant. The reconstruction helps us split longer FERs into shorter ones. Rule pruning will reduce the false positive rate in an ADS. We are mainly interested in daily network traffic, like the TCP dump. For example, in the following rule,

\[(\text{service} = \text{ftp, srcbyte} = 1,000) \rightarrow (\text{service} = \text{smtp, service} = \text{authentication}) \rightarrow (6)\]

is ineffective, because it can be reconstructed from the following two shorter rules:

\[(\text{service} = \text{ftp, srcbyte} = 1,000) \rightarrow (\text{service} = \text{smtp}) \rightarrow (7.\text{a})\]

\[(\text{service} = \text{smtp}) \rightarrow (\text{service} = \text{authentication}) \rightarrow (7.\text{b})\]

This reconstruction is more powerful if the window size is large. For smaller window sizes, the occurrence of an episode may be longer than the window size, violating the basic assumption in FER formation. This reconstruction may result in fewer false alarms.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Traffic Connections and Their Anomaly and Normality Scores (Dataset-I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ip_proto</td>
<td>src_ip</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
<tr>
<td>tcp</td>
<td>172.16.112.100</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
<tr>
<td>icmp</td>
<td>53.88.213.15</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
<tr>
<td>tcp</td>
<td>194.27.251.21</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
<tr>
<td>icmp</td>
<td>202.77.162.213</td>
</tr>
</tbody>
</table>

5 WEIGHTED SIGNATURE GENERATIONS FROM DETECTED ANOMALIES

In this section, we present a weighted signature generation algorithm to characterize anomalies detected by the ADS. The signature-based SNORT imports these
signatures and detects the same attacks subsequently. The traffic data set is a normal relation table consisting of N connections. A connection ci has M attribute-value pairs < aj; vi;j >, where 1 <= i<= N and 1 <= j<= M. The attributes are selected from the connection features and temporal statistics summarized in Table 1.

The ADS assigns an anomaly score and a normality score for each connection after processing a traffic data set. The anomaly score indicates the degree of anomaly that a connection is deviated from normal traffic, whereas the normality score indicates how close a connection is related to normal traffic. The ADS assigns the anomaly and normality scores to a given connection by comparing its FERs with the normal profile. In our ADS, the sum of anomaly and normality scores is normalized to be 1. Note that the sum is not necessarily a constant in other anomaly based detection systems. Table 2 shows an example of traffic connections and their anomaly and normality scores. Given a set of connections, as well as their anomaly and normality scores assigned by ADS, the problem is reduced to discover the most specific and discriminative patterns of abnormal connections. A connection pattern corresponds to a subset of < attribute; condition > pairs, which determine possible, attribute values in the episode. We define the anomaly or normality scores of a pattern as the sum of all anomaly or normality scores of all connections matching the pattern. We define signatures as those patterns that have high anomaly scores but relatively low normality scores.

5.1 Weighted Signature Generation

Let I – {i1; i2; . . . ; id} be a set of distinct items and T = {t1; t2; . . . ; tN} be a set of transactions. Each transaction ti consists of a subset of items from I. An item set X is a subset of I, and its support sup(X) is the fraction of connection containing X.

Fig. 5. Functional components in a weighted signature generation unit.
We find the frequent item sets whose support exceeds a user-specified threshold (\(\text{min\_sup}\)). We represent each connection as a transaction and each \(<\text{attribute}; \text{value}>\) pair as an item. The pattern of \(<\text{attribute}; \text{condition}>\) pairs are constructed from the frequent item sets.

Fig. 5 shows the functional components in the signature generation unit introduced in Fig. 2. First, similar abnormal connections are grouped together using a clustering analysis. Then, signatures are extracted for each group sharing some common characteristics. Attributes values of each connection are encoded into item numbers for mining the abnormal attribute-value pairs. After eliminating nondiscriminative item sets, the frequent item sets are decoded into \(<\text{attribute}; \text{condition}>\) pairs to form the anomaly signatures.

### 5.2 Cluster Analysis and Attribute Preprocessing

In traffic connections, different types of unknown attacks are often mixed together. The anomaly and normality scores assigned by ADSs are not sufficient to classify unknown attacks into different groups. It is difficult to discover accurate patterns or signatures for different types of attacks if they are mixed together. Different attacks have skewed distribution on connection volume. For example, different denial-of-service (DoS) attacks could have a large volume of short connections, whereas remote-to-local (R2L) or user-to-root (U2R) attacks could only have a few connections. Table 3 illustrates the attribute discretization and encoding in a Dataset-I. To separate different types of attacks, we use cluster analysis to group similar attacks together and generate signatures for each attack class.

We use the density-based clustering algorithm by Ester et al. [9] to obtain the partitioned cluster. Connections in low-density regions are classified as noises and thus omitted. Since each connection is characterized by symbolic and continuous values, we discretize the attribute values such as src_bytes in Table 3 by the entropy method introduced by Fayyad and Irani. We encode each pair \(<\text{attribute}; \text{value}>\) with an item number. Each item number has six digits. The first two digits represent the attribute index. The remaining digits are discrete symbolic values. For example, the pair \(<\text{proto}; \text{tcp}>\) could be encoded as \(3 \times 10000 + 2 = 030002\), where 03 is the index of attribute prot and 0002 is the index of the attribute value tcp.

Our signature generation is based on a weighted frequent item set mining (WFIM) framework. Algorithm 2 extends from this framework to generate anomaly scores. Let \(w_i\) be the weight of connection \(ti\), the weighted support of an item set \(X\) is defined by \(w\text{sup}(X) = \sum_{i\geq X, t_i\in T} w_i / \sum_{i\in T} w_i\). The purpose is to discover all item sets whose weighted supports are above the minimum support (\(\text{min}\_w\text{sup}\)). We call those item sets weighted frequent item sets (WFIs). The anomaly score weighs a connection. The \(\text{min}\_w\text{sup}\) helps select desired signatures.
Algorithm 2. Weighted a Priori Algorithm for Generating Signatures from Anomalies Detected

1: INPUT: A set of items I, a set of connections (that is transactions) T, weight wt of connection t, and minimum weighted support min_wsup
2: OUTPUT: Weighted frequent item sets X with wsup(X) > min_wsup
3: W = \( \sum_{t \in T} wti \);
4: k = 1;
5: L1 = \{i | i \in I \land \text{wsup}(i) > \text{min}_w\sup\}; \{ \text{Find all weighted frequent 1 item sets} \}
6: repeat
7: k = k + 1;
8: Ck = \text{apriori_gen}(k_1) \{ \text{Generate candidate item sets} \}
9: for each connection t \in T, do
10: Ct = \text{subset}(Ck; t); \{ \text{Candidates contained in } t \}
11: for each candidate item set c \in Ct, do
12: \text{c.weight} += wt; \{ \text{Add connection weight} \}
13: end for
14: end for
15: Lk = \{c | c \in Ck \land c.\text{weight} / W \geq \text{min}_w\sup\};
16: until Lk = 0;
17: return X = \bigcup Lk.

This algorithm is specified for weighted anomaly signature generation. If the support of an item set exceeds min_sup, then all its subsets must be supported. This principle is used by the a priori algorithm to effectively prune candidate item sets. We add items to item sets that are sufficiently large. We follow a weighted a priori principle: If the weighted support \( wsup(X) \) of an item set X exceeds min_wsup, all of its subsets have their weighted support exceeding min_wsup as well.

5.3 Signature Extraction and Mapping

For a given min_wsup threshold, the Weighted Frequent Item Set Mapping (WFIM) scheme discovers all WFI's whose weighted supports exceed the threshold. When we consider many features as discussed in Section 3.2, the number of discovered WFI's could be too large to construct the precise anomaly signatures. In particular, there are many redundant WFI's that have to be pruned before signatures can be generated. Instead of using all WFI's, we adopt a notion of maximal weighted frequent item sets (MWFIs) as a compact representation of all WFI's used. An MWF is defined as a weighted frequent item set for which none of its immediate supersets has a support above the min_wsup threshold. Since any subset of an MWF is also a WFI, the number of WFI's that an MWF represents will increase exponentially with the number of items in an MWF.

After the MWFIs of abnormal connections are discovered, we extract signatures by decoding the item numbers in MWFIs into <attribute; condition> pairs. The <attribute; condition> pairs represent the abstract signatures of the detected anomaly.
These abstract signatures are then mapped into specific signatures of a target IDS system such as SNORT or Bro.

**TABLE 3**
Mappings between Connection Attributes and SNORT Rule Keywords

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>SNORT Rule Keyword</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ip_proto</td>
<td>protocol</td>
<td>IP protocol type</td>
</tr>
<tr>
<td>src_ip</td>
<td>source IP address</td>
<td>source IP address</td>
</tr>
<tr>
<td>dst_ip</td>
<td>destination IP address</td>
<td>destination IP address</td>
</tr>
<tr>
<td>service</td>
<td>destination port No.</td>
<td>Service type, e.g. http</td>
</tr>
<tr>
<td>icmp_type</td>
<td>itype</td>
<td>ICMP message type</td>
</tr>
<tr>
<td>src_bytes</td>
<td>Dsize</td>
<td>Packet payload size</td>
</tr>
<tr>
<td>flags</td>
<td>Flags</td>
<td>TCP flags</td>
</tr>
<tr>
<td>land</td>
<td>Samcip</td>
<td>Source and destination addresses are the same</td>
</tr>
<tr>
<td>src_count &lt;n&gt;</td>
<td>threshold: track by src, count &lt;n&gt;</td>
<td>No. of connections to the same destination</td>
</tr>
<tr>
<td>dst_count &lt;n&gt;</td>
<td>threshold: track by dst, count &lt;n&gt;</td>
<td>No. of connections from the same source</td>
</tr>
</tbody>
</table>

Table 4 shows how connection attributes are mapped into the keywords in SNORT rules. Besides the compact representation of frequent item sets, we eliminate indiscriminative ones that appear very often in normal traffic. Since the min_sup threshold is relatively low to discover frequent item sets in normal traffic, the number could be very large. Recall that in the MWFI for Dataset-I, the <attribute; condition> pair is decoded as follows:

(ip_proto = icmp), (icmp_type = echo req), (1, 480 <= src_bytes < 1, 490); (dst_count > 10). ----> (8)

Using the attribute mappings in Table 4, we translate the signature into a SNORT rule as follows.

```
alert icmp$EXTERNAL_NET any <> $HOME_NET any (msg :“ possible pod attack”; itype : 8; dsize : 1, 480 <= 1, 490; threshold : type both, track by dst,count 10 seconds 1; sid : 900, 001; rev : 0; ). ----> (9)
```
6. CONCLUSIONS AND FURTHER RESEARCH

We summarize major contributions and make some suggestions for further work on automated detection of intrusions and anomalies in an open network environment. The co-operative IDS/ADS system applies to protect any networked systems, including LAN-based clusters or intranets, large-scale computational Grids, and peer-to-peer service networks, and so forth.

For further research, we suggest the following approaches for continued research and development effort. Distributed CIDS will advance state of the art in co-operative intrusion detection. Extensive benchmark experiments are needed on the DETER test bed for this purpose. Extending a centralized CIDS to a distributed one is highly recommended with strong collaboration over multiple IDS sites. This offers a logical solution to protect Grids, clusters, intranets, and so forth. Cyber trust negotiations and frequent alert information exchanges among distributed IDS sites are the key research issues yet to be solved.

7. REFERENCES

Data Privacy and Technical Solutions

Mr. Iqbal Khan, Department of Information Technology, Punjab, India,
iqbal7860@gmail.com

Abstract

The proposed paper attempts to address issues of personal privacy in a world of computerized databases and information networks from security technology to data protection regulation to Fourth Amendment law jurisprudence -- typically proceed from the perspective of controlling or preventing access to information and argued that this perspective has become inadequate and obsolete, overtaken by the ease of sharing and copying data and of aggregating and searching across multiple data bases, to reveal private information from public sources. To replace this obsolete framework, we propose that issues of privacy protection currently viewed in terms of data access be re-conceptualized in terms of data use. From a technology perspective, this requires supplementing legal and technical mechanisms for access control with new mechanisms for transparency and accountability of data use. In this paper, I present a technology infrastructure -- the Policy Aware Web – that supports transparent and accountable data use on the World Wide Web, and elements of a new legal and regulatory regime that supports privacy through provable accountability to usage rules rather than merely data access restrictions.

Objective of this document to record solution definition as advised by many corporate legal counsel and provide technological solutions to ensure the data protection and data privacy acts across multinational implementation of applications.

Keywords: Data Privacy, Database, Technical Solutions, Data Protection, Data, Privacy, World Wide Web, Data Privacy Act
Application of Cognitive Styles and Inference to Project Management

Dr. David J. Sudbeck, Washington University

Abstract

Project managers increasingly need better and more-integrated inference tools to handle their complex tasks. In particular, managers deal on a daily basis with uncertainty and inconsistency in their environment and in their facts at hand.

1 Introduction

Managers need better tools to handle uncertainty and inconsistency in project management and development. Traditional approaches have endeavored to quantify uncertainty in terms of metrics such as reliability (Sudbeck et al. 1989) while other approaches have endeavored to reduce uncertainty.

March and Simon (1958) noted the process of inference can be used to reduce uncertainty, and Simon won the Nobel prize in part for his work on bounded rationality (Simon 1960) and the need to infer ‘good-enough’ (and not necessarily optimal) solutions that ‘satisfice’ problems. In particular, Simon promoted the inference mode of ‘discovery’ or ‘abductive logic’ (Simon 1977) as a tool to find these solutions.

Along with the problem of uncertainty comes the problem of inconsistency. The groups involved in project planning have different viewpoints. For example, an analysis of a large corporation identified “12 different functional areas of business addressed by 26 different projects” (Henninger 1996).
Different viewpoints can stem from different dimensions of approach (i.e. leadership dimensions), from different dispositions of participants, and/or from different preferences of participants (thinking styles, etc.). Among these thinking styles are different uses of inference modes (see Figure 1).

A project manager must deal with these views and resultant inconsistency in the project plan. Approaches to inconsistency in planning have included a) preventing inconsistency (by fiat, by template, by threat, etc.) and b) resolving inconsistency as it is encountered.

Cognitive styles including approaches based on inference play an important role in what a project manager does and new tools are now available that project managers should use, such as progressing from old case-based management to newer inference-based computer-aided project management and agile methods.
2 Background

Project management is understood here as the managing of resources to complete projects according to constraints of time, cost, and quality (i.e. how well it meets project specifications).

A new agile version of project management is being promoted by IBM and other industry leaders. In the agile philosophy, there is less emphasis on pre-planning and more on flexible revision of specs by various agents of the team. As a noted agile author has said, the members of a team derive a specification by ‘scrapping it out’ (Ambler). Various researchers have begun to investigate impediments to applying these new methods in traditional development environments with differing management styles (Misra et. al.).

3 Model

Traditional project management has included inspection and review of plan specifications. Such an approach was fallible and computer tools were soon introduced to help with this task. These tools included tools for knowledge-based software development, case-based reasoning tools, and data mining and neural network approaches.

As Russo and Nuseibeh (2000, p. 14) explain

Specifications are assumed to be composed of many partial specifications (typically developed by different stakeholders), related to each other by means of pre-defined “consistency rules”. Each partial specification may or may not contain logical inconsistencies. However, the overall specification is defined to be inconsistent whenever at least one of the pre-defined rules is violated.

There are numerous variants of logic that have been enlisted to combat the problem of inconsistency. Three of the most popular variants are the traditional deductive and inductive logics, and the newer discovery or abductive logic.

Deduction can be thought of as traditional rules of the sort if $A$ then $B$. Deductive rules were used in early expert systems and today are reflected in sophisticated business rule engines.

Induction can be thought of deriving if $A$ then $B$ from numerous observations of if $a$ then $b$, if $a$ then $b$ etc., where $a$ and $b$ are individual instances leading to the general case if $A$ then $B$. Several contemporary business engines incorporate inductive reasoning.

Abduction can be thought of as the ‘reverse’ of deduction (and abduction is sometimes called retroduction). Given an object $B$ and if $A$ then $B$, we conclude $A$ is likely. An example if we observer wet-grass, and we know if rain then wet-grass, we ‘abduce’ or hypothesize that earlier we had rain. The $B$ is sometimes called the goal or observation,
while the $A$ is the explanation (if rain then wet-grass). Abductive systems for business and project management are just now starting to be employed.

### 3.1 Analysis (Including Detection of Inconsistencies)

Detection of inconsistency in a spec can be seen as detection of a violation of a rule such as “for a sentence $A$ including a system variable or entity, it cannot be the case that both $A$ and not-$A$ can be inferred from the spec” (Russo and Nuseibeh, p. 9).

There are several ways in which abductive reasoning can be used in this detection. One way is to make a goal of $P$ and not-$P$, where $P$ is a predicate or symbol in the specification. If an explanation can be abduced that satisfied this contradictory goal, we know the spec is inconsistent. Moreover, the explanations (technical term ‘abducibles’) found that lead to the contradictory goal indicates to us where the spec if inconsistent. On the other hand, if no explanation can be abduced for any such goal, the spec is consistent.

### 3.2 Revision

There are a number of researchers who point out “inconsistency is inevitable in real large-scale specifications” (Russo and Nuseibeh, p. 18). Abductive inference is applied by these researchers to either handle the inconsistency (i.e. remove objectionable inconsistencies while leaving inconsequential ones alone) or provide ways to re-establish consistency. The related science of complex event handling claims the “key to understanding events is knowing what caused them” (Luckham, p. 10).

An inference-based example of a specification could be for movement of a locomotive train (example adapted from Menzies). If an event $E$ has a train moving at time $t_2$ we can infer the train started at time $t_0$, was running at time $t_1$, and had no brake on at time $t_1$.

$$E(\text{moving}, t_2) \rightarrow E(\text{started}, t_0) \land E(\text{running}, t_1) \land \text{not-E(brake, t1)} \quad (I)$$

We posit a ground instantiation $EC(S) \rightarrow I$, with EventCondition at underlying time $S$. We then expect

$$EC(S) \land I(Sc) \rightarrow I(Sn)$$

Or, in other words, we expect ground-conditions for the specification plus the movement of the train at current time $Sc$ results in movement of the train at next time $Sn$. We then try to abduce a $D$ such that

$$EC(S) \land I(Sc) \land D \rightarrow \text{not-I(Sn)}$$
If we find such a D we know the spec is not consistent with the desired train-movement (i.e. perhaps the spec allows blizzard weather that halts the train; this ‘fact’ must be either resolved (i.e. spec for a snow-plow) or recognized as an allowed inconsistency).

If a D can be found to disprove the requirement, the D is an explicit indication of where the problem lies in the specification. The abduction approach thus has the advantage of being able to explain its conclusions.

3.3 Autonomous Agents

Work on social systems and agents in those systems have explored various ‘logics’ or ‘thinking styles’ identified in human agents. Abducible agents are proposed in (Kowalski and Sadri).

Such systems are particularly strong in temporal reasoning, where interaction of agents can quickly lead to conflicting requirements. These abductive systems recognize we cannot admit specifications that at the same point in time assert something and also deny that same thing. These systems can handle inconsistency across different times, recognizing a kind of evolution of requirements.

For example (Mello et al., p. 35), from the assertion

Y will accept a request within 2 to 10 minutes after the request is made

And the contradictory assertion

Y will not accept a request in the next 5 minutes of the request

An abduction system can infer the coherent

Y will accept a request within 6 to 10 minutes after the request is made.

These abduction systems can be made efficient enough to run in real-time through incremental operation (detecting inconsistencies as soon as possible), interleaving of assertions/retractions/abductions, and reduction of complex sentences to simpler more atomic sentences whenever possible.

4 Conclusion

Limitations of present tools for project management are increasingly being felt. In the realm of logics, abductive inference or the logic of discovery is finding increasing popularity.

An abductive approach, since it is based on logic, can work with just initial specifications, inferring inconsistencies. It can also quantify which inconsistencies are
significant, thus deciding which can be ignored. The well-known Bayesian net
approach was designed for abductive inference and is found in several vendor products.
(See Rodriguez et. al for applications to project management.)

Data mining and neural network approaches have the disadvantage of not being able to
explain why they make their conclusions. A limitation of case-based reasoning (which is
sometimes called rule-induction) is it assumes past-cases are available for analysis.
Satoh (1996) has shown how case-based reasoning can be translated into a subset of
abductive logic programming.

Several efforts have been recently devoted to extending logic programming to perform
abductive reasoning. Abduction is a form of reasoning which allows one to compute
explanations for observations. Moreover, it is a form of non-monotonic reasoning, since
explanations which are consistent in a given context may become inconsistent when
additional information is added. In fact, it is well-known that abduction provides an
alternative formation of default reasoning.” (Giordano et. al., p. 1-2)

5 References


Handbook of Logic in Artificial Intelligence and Logic Programming(D.M. Gabbay, C.J. Hogger, and J.A.
Robinson, Eds.) Oxford University Press.

Mathematics and Artificial Intelligence.

Luckham, David. 2002. The Power of Events: An Introduction to Complex Event Processing in
Distributed Enterprise Systems


Mello P., P. Torroni, M. Gavanelli2, M. Alberti, A. Ciampolini, M. Milano,
interaction amongst computees.

Journal of Human Computer Studies, 45, p. 305-35

Misra, S. et. al. 2006. The Organizational Changes Required and the Challenges Involved in Adopting
Agile Methodologies in Traditional Software Development


Early Muslim Scholars and Binary Mathematics

Numan Tugan
Milwaukee, Wisconsin, U.S.A.
numantugan@yahoo.com

&

Mohammad Amin & Pradip Peter Dey
National University, San Diego, California, 92127, U.S.A.
mamin@nu.edu ; pdey@nu.edu

ABSTRACT:

Using the concept of dualism, early Muslim scholars envisioned the world as an expression of events in terms of two irreducible opposite elements. Whereas others saw this opposition as a source of conflict, these early Muslim scholars saw it as a source of cooperation and interaction, which led to the development of 0-1 mathematics. Historical evidence of their work, examples of their achievements, the thinking set and mathematical operations on this set will be presented. Illustrations of tables and calculations are included.

Keywords: Dualism, 0-1 Mathematics, Mathematical Thinking Set.

HISTORY

The concept of dualism is mentioned in the Quran, the holy book of Muslims, in many places. The following verse, “And He it is who has created all opposites (or pairs)…” [1], and similar verses clearly indicate that the universe is made of opposites and pairs, and is a place for humans to think, contemplate and learn lessons from these opposites. Contemplating these types of statements, Muslim scholars have expressed many different interpretations regarding this. Some commentators regard the term “pairs” as synonymous in this context with "kinds" [2]. Others see in it a reference to the polarity evident in all creation. Some say that it denotes the concept of opposites in general, as "sweet and sour, or white and black, or male and female,” that everything in creation has
its complement, "like high and low, right and left, front and back, past and future" [3].

The expression of these different views or opinions started with the earliest Muslim scholars of the seventh and eighth centuries, and continues even today. Discussion of their various opinions on the subject of dualism helped them to develop a new form of logical thinking process called “kalam.” A rigorous scientific approach to this study allowed the development of a methodology for the description and analysis of creation, nature and human existence. Many valuable books were produced during this period, and these early scholars were succeeded by their equally erudite students [4].

After the first generation of Muslim scholars, the second generation and beyond undertook philosophical studies and translated the works of the early Greek philosophers, whose opinions on dualism varied considerably from their own. They came to consider the two irreducible opposite elements of dual pairs as sources of conflict with each other. One gains superiority over the other through conflict and defeat [5]. Those scholars who regarded philosophy as a foreign ideology and rejected its principles did not follow its methods, but used purely Quranic thinking methods. They understood opposites as the basic structure of things, a source of cooperation and interaction. They identified opposing forces not as enemies but as cooperating pairs, helping and completing each other. This type of thinking process helped them to understand the laws of nature and utilize them for their benefit [6].

During the 8th and 9th centuries, those scholars who followed the path of the Greek philosophers were active in politics and controlled all the education and the decision-making mechanisms of the “House of Power” of Caliph in Baghdad and in Cordoba. The other scholars were persecuted and dispersed all over the empire. They migrated to many different places, especially to Samarkand, Bukhara and Khwarizm, to the east. But the scholarly debate never ceased between these two groups. The 10th century produced a tremendous number of scholars, scientists and mathematicians [7]. In this century, a Muslim scholar named Muammad ibn Mūsā al-Khwārizmī introduced zero as a number to the study of mathematics, as well as the methodical problem-solving technique called algorithm, which is derived from his name [8]. Territories such as India captured during this period of the 10th century were influenced by these scholars in science and mathematics [9].

This momentum of scholarly studies produced many scientist and academics throughout the centuries, which formed the basis for many of today's discoveries and theories. A few examples of these are the first mechanical robot of Al Jazary[10], the map of Ottoman Naval Captain Piri Reis[11], and the 17th century flying apparatus of Ottoman aviator Hazerfen Celeby[12], who flew with artificial wings more than 4.5 miles. One contemporary scholar even described the atom as a type of small motor or machine [14]. Many extant works of these scholars are collecting dust in libraries throughout the world, like treasures waiting to be rediscovered. Some libraries, such as the Suleimaniye Library in Istanbul, have undertaken massive projects to translate this kind of material and make it available to the public. Other valuable old books have not been read in centuries, and might never be seen again.
Our focus in this paper is (RML), another example of the early Muslim scholars' works upon which binomial mathematics is based. The concept of dualism is represented in 0-1 terminology to describe one aspect of the thinking process, which is decision-making. Muslim scholars were aware of the thinking process and they refer to it many times in their writings. For example, Said Nursi refers to the speed of thought relative to the speed of other phenomena, saying, “There are many degrees of speed, such as the speed of sound, the speed of electricity, the speed of light and the speed of thinking or imagination.” [15] The thinking process was treated like any other natural phenomenon by the Muslim scholars. They described this phenomenon through logical expressions while explaining the concept which involves thinking and imagination in the terminology of 0-1 [16-20]. The concept of dualism is deeply embedded in these writings, one part of the framework of their way of looking at the world.

These scholars described the information which goes into the human brain in terms of small discrete entities of information or dots called Nuqat (singular nuqda). Information is stored in the brain also as discrete “dots.” When the information is recalled, it is in the same format, one dot at a time, consecutively. There is no simultaneous process, such that two different pieces of information or dots could go in or out at the same time. The speed of this process is what is referred to above as the speed of thinking or imagining. The information dots are so small and so close to each other, that it might appear that the dots are being processed simultaneously.

In the decision-making process, the stored information must be recalled and acted upon, again as a series of discrete dots. Because it would be impossible to consciously examine the dots of information individually, each dot is represented by a randomly generated symbol which is used for comparison to the symbol following it. A mathematical formulation to assist in the decision-making process is as follows:

The set of Nuqat is expressed by \((\mathbb{N}, \oplus)\), all \(a\) and \(b\) are members of \(\mathbb{N}\), such that there is an operation on this set expressed by \(\oplus\) (=dot operation) so that

\[
a \oplus b = \begin{cases} 
1 & \text{if } a = b, \text{ and } 0 \text{ if } a \neq b 
\end{cases},
\]

(= expressed as same as)

If two dots of information coming out consecutively are the same symbol, then the result is 1. If two consecutive dots are not the the same, then they are opposite, and will negate each other, and then the result is 0. A result of zero is equated with a decision of yes, a result of one is a no. I will present a table of calculations as an example.
Muslim scholars have produced various books expressing the above ideas in different terms and interpretations. [21-24] These scholars approached the subject in two different ways, one of which is deterministic and the other probabilistic. Events in the decision-making process are classified as either deterministic (predictable) or probabilistic (random). There is very little information available on the deterministic concept [25]. Most of the books currently available were written on the probabilistic concept, which is somewhat different from today's study of probability. Further development of these concepts, combined with aspects of Raml, a variant of geomancy, are referred to in many subsequent writings. [21-24].

The study of these concepts could be applied to the areas of artificial intelligence, neuroscience, nanoscience, and robotics.

REFERENCES:
[1] - Quran, chapter (43:12)
[5] - Wāsil Bin Ata
[6] - Sayyid Qutb, In the shades of the Qur'an

[11]- Piri Reis

[12]- Hezârfen Ahmed


[14]- Said Nursî

[15]- “The Words, Thirty-First Word, speed of sound, electricity light, and thinking (imagination)…” Said Nursî

[16]- Muhammad ibn Muhammad ibn Mahmud Abu Mansur al-Samarqandi al-Maturidi al-Hanafi (d. 333 AH / 944) Kitab al-Maqalat

[17]- Jalâl al-Dîn Rûmî (1207-1273) and

[18]- Muhyî al-Dîn Ibn ‘Arabî (1165-1240),


[20]- Risale-i Hesab / Yegani Sinan Çelebi. Suleymaniye Library, Istanbul

[21]- Risale-i ilm-i remil , Suleymaniye Library, Istanbul

[22]- Risale-i Cifr-i Cami / Hüseyin el-Ahlati, Suleymaniye Library, Istanbul

[23]- Risaletün fi Ebvabi'r-Remil / Cemaleddin Ömer b. Osman ez-Zennati ,Suleymaniye Library, Istanbul

[24]- Fenn Remil, Suleymaniye Library, Istanbul

[25]- İbtalü Temassi'l-Küre li-Nukat Muhtelife fi Zeman V. / Molla Kasim el-Germiyani el-izari, Suleymaniye Library, Istanbul
Acceptance Test Driven Story Card Development for XP (Agile Software Development)

Chetankumar Patel
School of computing
Software engineering research group
Leeds Metropolitan University
Beckett Park, Headingly
Leeds, LS6 3QS, UK

Muthu Ramachandran
School of computing
Software engineering research group
Leeds Metropolitan University
Beckett Park, Headingly
Leeds, LS6 3QS, UK

Abstract
Requirement elicitation is a process to collect information from the end users or stakeholders of a system. In traditional software development models, requirements are predefined and sent to the developers to do analysis from various stakeholders. Requirement elicitation and quality related to requirements are the challenging issues in the software development life cycle. XP recommends an on-site customer to represent their requirements through user stories on story cards. Generally, customers have rarely a general picture of the requirements or system in their mind which leads to problems related to requirements like conflicts, missing requirements, and ambiguous requirements, etc., and not addressing non-functional requirements from exploration phase. Two-thirds of the projects are failed because of ambiguous and incomplete user requirements, and poor quality of the requirements. Acceptance testing is also one of the key issues, which is often left unsolved or uncovered to verify user requirements in XP (Agile software development methodology). To solve these problems, we propose a New Improved Requirement Elicitation (Gathering) Process framework for story cards in XP to improve the quality of user stories to capture functional and non-functional user requirements on story cards. Our approach involves adopting acceptance tests right from story cards as soon as requirements are clearly understood. We also aim to demonstrate how acceptance tests will improve story card-driven agile software development (XP). We believe this will enhance the popularity of Agile-based software development.

Keywords
Extreme Programming, Requirement Engineering, Acceptance Test, User Story, Story Card

Introduction

User requirements elicitation is one of the early stages of the software project development which records the description of the system how it has to work or what system has to do or statement of how it should do it (Somerville and Kotenya 2000). Requirements elicitation process is one of the challenging processes in the software development methods. In traditional software development methods, end users or stakeholders predefined their requirements and sent to the development team to do analysis.
and negotiation to produce requirement specification. In traditional software development methodology it is expensive to deal with requirement change after careful analysis and negotiation. This problem is well tackled by the XP, which is one of the agile software development methodologies.

Extreme programming is a lightweight and incremental software development methodology which directly involves their customers and developers. XP is actually consisting of 12 practices based on the four values (Beck 2000). Such as test driven development, continuous integration and refactoring can be practiced by individual developer because they directly gives benefit to the developer while some of them are hard to implement individually like Planning game, acceptance testing, and small releases because they require buy in and participation of people from the team or from outside (Rogers 2004).

In XP, Development starts with planning game where customer writes user stories on story cards. Those cards are estimated by the developer, based on those estimation customer priorities them depends on their needs to establish a timebox of an iteration. Developers develop those story cards through pair programming and test driven development. At last customer provides acceptance test to accept the developed functionality. In between they consider all of the XP practices in mind to improve the quality of the software.

Story cards are playing vital role in XP. It is the unit of functionality and prioritization. Story cards are one of the important aspects of the XP. Story cards are written by the customer in XP to articulate their business needs. According to Cohn story cards must be testable, estimatable, valuable to the customer, small and independent (Cohn 2004). These story cards must be written by the customer because they know their business need very well compared to developer. Consider the figure 1 which shows a traditional story card as proposed in (Beck 2000).

![Figure 1 Traditional Story Card Used in XP Projects](image-url)
XP strongly recommend an onsite customer to write their business need. Business is well understood by the customer. Generally customers have rarely a general picture of the requirements or system in their mind (Kotyana and Sommerville 1997). Different stakeholders have different needs. Often an onsite customer is not complete and precise during story cards writing process and do not predict their working system without seen it, which will lead problems related to requirements like requirements conflicts, missing requirements, and ambiguous requirements etc, and not address non-functional requirements from exploration phase. Incomplete requirements, requirements conflicts, missing requirements, and ambiguous requirements are going to be a defect on working system. Two third of the projects are failed because of ambiguous and incomplete user requirements, and poor quality of the requirements. In most of cases non-functional requirements are not covered in the exploration phase. Non-functional requirements play a significant role with functional requirements in software development methods. Non-functional requirements improve the quality of the software. If they are not considered in early stage then it is difficult to address them in final product. In XP there is a question related to quality is unanswered. There is still scope to improve story cards.

Acceptance tests are the test written by the customer to evaluate that the developed functionality works according to their requirements or business needs. Acceptance testing is one of the challenging issues in the software development lifecycle as it is often neglected on an XP projects. It is perceived as just being to hard to get right (Rogers, R.O. 2004), Which causes a problem related to requirement quality and become difficult to check weather system does what it’s intend to do it or not. Massive development efforts based on unproven business requirements are a waste of time effort and money (Bodell, R. 2003). The CHOAS report published in 1995 shows that almost half of the cancelled projects failed due to a lack of requirements engineering effort and that a similar percentage ascribes good requirements engineering as the main reason for project success([Chaos 95]

As a result of this investigation we propose a new prototype to improve requirement elicitation process in XP and new improved framework of story cards to write user stories. This will help to customer and developer to improve the quality of the user stories or story cards, and to address functional and non-functional requirements on story cards. In this paper we discuss about extreme programming and requirement elicitation process through story cards first and then after the challenges and problems on XP software development methodology. This is followed by discussion of related research regarding to an improvement framework of the story cards for XP based projects.

**A New Improved Requirement Elicitation (Gathering) Process in XP**

Extreme programming relies on customer interaction to identify or to tell which features to implement in next release. Stories are written by customers perhaps with the help of agile development member. Therefore each user stories and story cards must be split
into a unique and independent requirement. In many cases a process to get user-visible requirements is hard to impossible. They do not supply any infrastructure requirements.

As a part of this investigation, in our approach we recommended a customer or stakeholder who is on site has comprehensive application domain and business knowledge to put a developer into the picture.

Application domain knowledge and customer business knowledge from customer will help developer to focus on stakeholders’ business needs and requirements and help them to cover any missing functionality. However it is a question of how much risk is involved by not spending sufficient time selecting the right customers and prioritizing them and their respective requirements. It is impossible to have all requirements just from one person. In the following Figure 2 we propose a new improved requirements elicitation process in XP, which emphasizes on customer and developers communication on customer’s domain and business knowledge. XP recommend that developer and customer discuss about their proposed system. We assume that the onsite customer is a domain expert or he understood their business requirement very well and he or she is able to explain their business needs at a glance to Developers to put them into picture of system. Customer business knowledge helps developers to understand how the system will affect and interact with the different part of the business, and help to identify and understand different stakeholders, who are directly or indirectly affected by the system.

Stakeholders identification is one of the important and challenging parts of the requirement elicitation process in XP (Agile software development) and traditional software development. It is quite difficult and risky to get information or collect requirement for whole system from the single prospective. So we come with the idea of identify all stakeholders and their needs requirements to build a good system. XP most likely assume that an onsite customer already have a wide view of all requirements of the system but in most cases it is not and in this case XP also struggling to find out all stakeholders requirement form one person and priories them according to that persons limited knowledge of requirements. Here developers going to ask a series of question to identify stakeholders of the system who are going to affected directly or indirectly by the system.
At the end of successful discussion between customer and developers, customer starts story elicitation process and writes the draft statement of requirements on story cards. These story cards are further analyzed, which assist to customer and developer to identify any problems and missing functionality. Missing functionalities become defect in working software. This scenario will help developers to focus on non-functional requirements. Also helps to keep in mind what they have to do to improve all the aspects of the business through structured system. Identified problems are being negotiated between developers and customer to acquire story cards.

**Acceptance Test Driven Framework of Story Cards**

We introduce acceptance tests directly on the story cards. Acceptance tests on exploration phase for two reasons 1) to identify the verifiable requirements from different stakeholders and 2) This will help developer in effort estimation. Following figure 3 shows new template of the story cards.
There are certain advantages of the acceptance test to introduce directly on the story cards. Acceptance test do the validation of the requirements, helps to write the user stories on the story cards on a verifiable form instead of vague, ambiguous or best guess requirements. Acceptance test also improve the estimation of the story cards as well. If story cards contain too many acceptance test than that means story card is big to fit in iteration and to do estimation. So acceptance test on story cards give an idea to split stories into a smaller stories to fit in iteration, and smaller stories wit acceptance test have a perfect estimation. Acceptance test directly tell to customer what they have to develop instead of guess.

We introduce a new section called point to consider on the story cards. This will helps to identify the goal of the users and other non-functional requirement related to that story card. This is one kind of the constraint where developers and end user or customer add their requirements which are directly or indirectly apply to the story cards or entire set of story cards of whole software. We apply this story card to the real world user story and explained how it improves the requirement or user story on story cards. Figure 4 shows the story cards written through a traditional way, which is just a short and vague statement of user requirement.
This traditional story card does not provide enough information. It is really difficult for the customer to estimate and test the user requirement as well. Also this is not come with any kind of the user acceptance test as well. Acceptance testing is remain unsolved in most of the projects at moment, which is need to be consider.

<table>
<thead>
<tr>
<th>STORY CARD NO:</th>
<th>Project Name</th>
<th>E-Commerce</th>
<th>Estimation: 4 Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Story Name:** User Registration

**User needs to register with unique username and password before purchasing anything from the online store**

**Acceptance Test**

1. User Id must be unique
2. Try to register with duplicate user id and password
3. Try to register user name only
4. Try to register with password only
5. Forget Password Link

**Note:**

User can view or visit store as a Visitor but needs to register before purchasing anything

**Risk:** Low

**Points to be Consider:**

There isn’t any non-functional requirement at this stage

---

**Figure 4 User Story captured on a Traditional Story Card**

**Figure 5 User story captured through the improved requirement elicitation process by using the new acceptance test driven story card template**
The figure 5 shows the same user story captured through the improved requirement elicitation process by using the new acceptance test driven story card template, which captures user requirement precise, correct, testable or verifiable and enough sort of information with acceptance test. This will help us to estimate human efforts to convert user story cards into working code through the relationship of acceptance tests and effort estimation.

**Conclusion**

Our previous work on knowledge base support on story cards, which provides guidelines to capture requirements on story cards. In this work we introduced an improved framework of story cards which helps developer and customer to address all functional and non-functional requirements of the software. This will improve the story cards and story card based agile software development. We also focus here to write acceptance test with story cards instead of during the iteration planning.

**Future Work**

This work is just a prototype and we going to adopt this concept on to real time projects. We are also working on guideline based story card development and value oriented prioritization matrix to priories story cards and customer priority.

**References**


CHAOS, Software development report by the Standish group 1995).


Analysis of Near-infrared Phase Effect on Biometric Iris Data Demonstrates Viability of Iris-scan Biometric Device for Identity Authentication in Public Venues

Brady R. Stevenson and Gordon W. Romney
National University, San Diego, CA

Abstract
Correctly and quickly validating the identity of individuals to permit entry to a venue is an increasingly challenging issue in modern society. One common objective is to allow entry to authorized individuals and to restrain all others. Transportation and border venues, that process a high percentage of human movement, are especially critical and vulnerable from a national safety perspective. In this context, the selection of biometric technology that uniquely identifies a human and meets specific operational criteria is fundamental. Finding a statistically reproducible process, quick to capture data, non-invasive to humans, and which functions in a variety of ambient environments becomes the challenge.

Iris-scan technology was selected as the candidate identity authentication mechanism of this research as it is quick to capture data, is non-invasive and produces statistically verifiable data. Prior research had left unresolved the issue of whether normal, ambient environments with variable levels of fluorescent lighting such as airports, rail stations and passport control kiosks might produce unreliable data due to wavelength sensitivity of iris cameras. Prior studies of iris data variations from infrared wave interference of halogen, incandescent, and sunlight were inconclusive relative to near-infrared wavelengths between 700 and 900 nm. Such wavelengths are produced in fluorescent light sources, frequently found in the targeted venues, that commonly emit 30% of their heat in the near-infrared spectrum. The experiment utilized a Panasonic BM-ET300 iris camera scanning nine hundred fifty-four (954) subjects with a variety of ethnic eye configurations using variable background light of a normal fluorescent lamp. Results showed increased heat from fluorescent illumination has no statistical significance due to the near-infrared waves from the camera and the fluorescent lamp. The results of the test confirm that data loss will not occur as heat is increased in the fluorescent light source. Hence, image corruption which would produce unreliable authentication data would not be caused by increased levels of fluorescent light conditions. Use of this particular iris camera in normal ambient lighting conditions should prove viable for biometric identity authentication purposes.
Challenges in Vendor Collaboration in A Mix Mode Offshore Centre - Coexistence Of Captive And Vendor Offshore

Satyen Vats & Rathi Dasgupta

Abstract
Companies with head quarter in US and Europe more and more are using Offshoring to leverage cost and augment skilled staff. Move to create a wholly owned captive centre is considered to be a smart move to save cost and create reusable competency. The most successful model is the ecosystem of coexistence of captive center and vendor partners to do projects. Due to increasing competition in resourcing companies are using vendors to do staff augmentation for spiked requirement during project life cycle. The current research focuses on the challenges in vendor management and the best practices model.

1. Introduction
Non-availability of knowledgeable resources, reduction of fixed cost, ever increasing cost of wages and attrition are some of the challenges, faced by captive centers [4], which are addressed through engagement of offshore vendors. This engagement promises best of both worlds in terms of flexibility of vendors [1] and control of a captive center. But it also creates a complex set of problems that require a carefully thought, planned and focused approach towards vendor management [2] [3].

The vendor management challenges involve the following root causes:

1. Business objectives –
   a. Engagement of offshore IT vendors with offshore captive centers involves the problem of conflicting business objectives.

2. Ownership and expectations –
   a. Conflicting business objectives for different entities engaged in a typical captive center project can easily lead to, in absence of a universally agreed and accepted project charter, conflicting/different expectations and non-uniform understanding of ownership of various aspects of a project, on part of all entities.

3. Development processes and methodologies –
   a. Almost always, vendor’s development processes and methodologies are different from captive or client’s internal development processes. Absence
of a common process framework during the execution of a project leads to significant cost increase and delays on projects.

4. Corporate culture –
   a. Differences in corporate culture present serious obstacles to working relations between the vendor and captive organization –
      i. Organizational structural and reporting
      ii. Organizational values and behavioral norms
      iii. Management approach –
           1. Decision making
           2. Risk appetite
           3. Monitor and control mechanism
      iv. Employee evaluation and skill level assessment

5. Knowledge transfer –
   a. Knowledge transfer from vendor resources to captive center resources is considered to be an important success factor for a project as well as long term success of the captive center. Multiple

6. Skill level –
   a. Domain knowledge
   b. System understanding
   c. Technology competency

7. Lack of team camaraderie

2. Conflict of business objectives

   Engagement of offshore IT vendors with offshore captive centers involves the problem of conflicting business objectives. It is perceived that captive centers and vendors are competitors and, in fact, this perception is true to certain extent. Management of parent organization does not share and discuss the vision, objectives and long term plans associated with the captive center. Vendors, in the absence of any clarity regarding their future relationship with the parent organization and captive center, become skeptical of the captive center.

   i. Lack of clear accountability
   ii. Captive center is considered as threat to business
       - Vendors create nexus with parent organization's employees who are against captive center
       - Miss-representation of facts to highlight captive center as incompetent
   iii. Lack of support for captive center's growth from vendor
       - Resistance towards sharing of knowledgeable resources
       - Resistance towards working in a partnership as a single team
   iv. Vendor and captive center compete against each other
- In-effective communication
- Lack of co-operation among project teams
- Lack of camaraderie among team members
- Blame game

3. Ownership and expectations

Conflicting business objectives for different entities engaged in a typical captive center project can easily lead to, in absence of a universally agreed and accepted project charter, conflicting/different expectations and non-uniform understanding of ownership of various aspects of a project, on part of all entities.

4. Different processes and methodologies

Vendor’s development processes and methodologies are different from captive or client’s internal development processes. Absence of a common process framework during the execution of a project leads to significant cost increase and delays on projects.

i Different software development life-cycle methodology and framework
  - One organization may prefer RUP like approach of exact specification and detailed documentation while other may prefer agile methodologies.

ii Different project management approaches
  - Vendors adopt client's management practices while working at client location but their offshore project management practices are very different.

iii Different level of process maturity
  - Vendor may be at quality certification of CMM Level 5 where as the captive center may be at Level 2.

iv Different quality assurance processes
  - Application of different quality assurance approaches to a single application-code leads to in-consistent quality of code.

v Different code libraries, development and testing tools and version maintenance and documentation systems
  - Developers and testers are used to certain libraries, tool and systems. Use of multiple tools and systems, for same set of activities, on a project leads to problems of code integration and knowledge management towards later stages of the project and during maintenance and support activities.

5. Corporate culture differences

Differences in corporate culture present serious obstacles to working relations between the vendor and captive organization.

i Organizational structural and reporting

ii Management approach –
- Decision making
- Risk appetite
- Monitor and control mechanism

iii Employee behavioral norms and expectations
- Education and dedication are great strengths of offshore, particularly Indian, staff but they lack people skills required for a multi-national and multi-cultural work environment. These people skills are at the root of ineffective communication and mis-match of expectations at the team level.

iv Employee assessment and rewards and recognition
- Performance assessment and rewards and recognition are an important aspect of employee motivation. Captive centers follow parent organizations assessment and recognition policies that are very different from vendors’ practices. Presence of non-uniform assessment and recognition practices and different assessors as well, for team members of the same team also contributes towards lack of team camaraderie.

v Skill evaluation and project allocation
- Domain knowledge
- Technology competency
- System understanding

6. Knowledge transfer
Knowledge transfer from vendor resources to captive center resources is considered to be an important success factor for a project as well as long term success of the captive center.

7. Conclusion & Recommendation

Price is a key differentiator, in evaluation of vendors, for a short term relationship and small piece of work. But outsourcing has moved beyond commodity-based, cost-driven engagements and saved-dollars is not the only value-add provided by offshore vendors. To derive additional benefits from outsourcing, customers need to forge a long-term strategic partnership with the vendors. Even though customers are highly focused in their selection process of vendors but they often miss this relationship aspect of vendor management. Many vendors are large successful companies and they look for a win-win relationship with their customers. Whenever customers take the conventional approach of squeezing the maximum out of vendors and vendors are not able to make their margins from an engagement, they loose focus from such engagements and make other profitable engagements as higher priorities.

Long term partnership with vendors –

Defining space of operations for customer, captive center and vendor, and sharing of long term goals and expectations are the key steps required in establishing a
strategic partnership with the vendors.

i. Customer and vendor need to create blueprint of a win-win relationship, with clearly defined role for the vendor, at early stages of their engagement.

ii. Price bench-marking, SLA compliance and contractual aspect of engagement should be replaced with business value, change management, knowledge transfer, domain knowledge, and process maturity.

iii. Create and maintain a system of accountability on all participants

iv. Vendor relationship should be maintained at all levels.

v. Establish constant dialogue to better understand each others business, business objectives, culture to facilitate value addition and better solutions.

vi. As vendor engagements are changing from SLA based contracts to strategic partnerships, customers and vendors need to innovate on pricing of services. A revenue based payment model can bring greater amount of ownership and accountability from the vendor.

Uniform processes and methodologies:

Customer and vendor teams should evaluate the differences among their processes and agree on a common development framework, project management approach and quality assurance methodology along with common code libraries, version control system and development and testing tools.

i. A governance body should be created to monitor and control progress of projects.

ii. Well defined ownership should be defined and compliance should be ensured.

iii. Tollgate process should be established to ensure completion of each phase before moving to next phase.

iv. Milestones should be clearly defined along with details of deliverables and completion dates.

v. An air-tight communication plan and reporting mechanism should be established to ensure early identification of issues and their quick resolution.
References

1. Vendor Management Tips – Building Relationship
   (http://www.computerworld.com/managementtopics/management/story/0,10801,99816,00.htm)
2. Vendor Management (http://cio.ittoolbox.com/topics/t.asp?i=426&p=426&h1=426)
3. A new way to manage vendors
   (http://www.cio.com/article/2034/A_New_Way_To_Manage_Vendors)
4. Dynamics of Captive IT centers, Rathi Dasgupta, Proceeding of 6th International Conference
   on Smart Sourcing, Atlantic City, 2008
COMPARATIVE STUDY AND ANALYSIS OF FLANK WEAR OF SINGLE POINT COATED AND UNCOATED TURNING TOOL INSERTS USING ACOUSTIC EMISSION SIGNAL PROCESSING AND NEURAL NEWORK TECHNIQUES

Mr. N.PRASANNA,
Assistant Professor- Department of Industrial and Production Engineering,
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, INDIA
E-mail: nprasa52@yahoo.com

Mr. SRINIDHILYR,
Assistant Professor- Department of Mechanical Engineering,
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, INDIA
E-mail: shreenedhe@gmail.com

Dr. M.SUKUMAR,
Professor & Head- Department of Instrumentation Technology,
Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, INDIA

Dr. C.S.VENKATESHA,
Professor-Department of Mechanical Engineering
University B.D.T.College of Engineering, Davangere, Karnataka, INDIA

ABSTRACT

Wear of cutting tools is very important and various output generated by the study and analysis of each tool is extremely useful in the tool condition monitoring process in general and in making efforts to determine the expected tool life in particular. Wear of the tool tip generates poor surface finish and an unexpected tool failure may damage the tool, work piece and sometimes the machine tool itself. Advanced manufacturing demands an optimal machining process with proper control of machining parameters. Many problems that affect optimization are related to the diminished machine performance caused by worn out tools. One of the most promising tool wear analysis and tool condition monitoring techniques is based on the acoustic emission (AE) signals. The generation of the AE signals directly in the cutting zone makes them very sensitive to changes in the cutting process. Experiments have been conducted on EN-31, EN-8, EN2A and Mild Steel specimens using Uncoated and Coated carbide inserts (Kennametal, India make) on a high speed lathe for the most appropriate cutting conditions. The AE signal analysis (considering signal parameters such as, ring down count (RDC), rise time (RTT), event duration (EDT), energy (ENT) and peak amplitude (PA).

Key Words: AE, RCT, RTT, ENT, EDT, PA

1 INTRODUCTION

Reliable monitoring of machining operations facilitates economical uses of modern, capital intensive, minimally manned production systems. Tool wear, chipping or catastrophic tool failures are the major obstacles in producing high quality products. Research during the past few years has led to different methods of monitoring either directly or indirectly the state of single-
point. However, direct monitoring has not reached industrial applications because these methods interrupt the machining process, time consuming and inefficient. Indirect monitoring methods for detecting signals related to cutting tool state have become successful in recent studies. Acoustic Emission (AE) methods, monitoring stress wave propagation’s released by material undergoing deformation or fractures in the cutting zone, are one of the most promising tool monitoring techniques.

2 RESEARCH WORK

The purpose of research work was to monitor tool wear (Flank wear) by direct method and to compare and correlate the results from other indirect method and training of neural network. Artificial neural network toolbox under MATLAB V7 was used to develop the program. Experiments using single point Coated Carbide inserts were carried out on En 2A, En8, EN-31 and Mild Steel (MS) specimens. The depth of cut was fixed at 0.5 mm at two speeds namely 500 rpm (94 m/min) and 1000 rpm (188 m/min) at feeds of 0.07 and 0.09 mm/rev respectively. Various parameters like acoustic emission, cutting forces, temperature, surface roughness, vibration are known to have an affect on degree of tool wear during cutting. Acoustic emission and force signals were selected as sensing fields. Since sound signals emitted during turning process are of a very high frequency, relatively distinct signal with almost zero is obtained. Acoustic emission sensors were mounted near the tool to collect acoustic emission signals released during cutting, which is passed through a pre-amplifier and then converted into electric signal and is collected in a computer. The flank wear measured are tabulated in the wear data sheet. The AE signals readings were taken for around 20 trials for which, considerable wear were observed during preliminary work. The AE parameters and flank wear are tabulated...

![Experimental Set Up](image)

Experimental work is shown in the figure. AE sensor is placed on the tool holder and connected with AET-5500 (Acoustic Emission Analyzer) through the pre amplifier and the AET-5500 main frame connected with the host computer for data acquisition and for online display of the data. Frequencies that are commonly used for AE testing lie between 50 kHz and 2 MHz. Various acoustic emission parameters were considered for analysis which are inputs to train the neural
network. After training the network, the weights obtained were saved. These weights are then used to compute the flank wear for a given trial. This program can be used to control the turning operations by deciding values for tool wear depending on cutting conditions and its application by experienced personnel.

3 ACOUSTIC EMISSION TECHNIQUES - FOR MONITORING TOOL WEAR

The acoustic emission technique (AET) is a nondestructive evaluation [9,10] and has shows the possibility for characterization and wear assessment in machining operations using conventional and non-conventional materials. This method has also been widely used to detect process changes like tool wear in metal cutting. In-process tool wear monitoring has presently acquired more importance because of flexibility and provides scope for the optimal utilization. The AE response from metal cutting changes as soon as the cutting tool starts the first cut and continues to change as the cutting tool wears. The major advantage of using AE to detect the condition of tool wear is that the frequency range of the AE signal is much higher than that of the machine vibrations and other interferences and relatively uncontaminated signal can be easily obtained by the use of a high pass filter. In addition, AE can be measured by simply mounting a piezoelectric transducer on the tool holder without being interrupted during the cutting operation.

4 MEASUREMENTS AND PREDICTION OF TOOL WEAR

Many types of tool wear are practically seen and some of them are measurable. Prominent amongst them are flank wear and crater wear. Tool wear [1] starts immediately when the new tool is put into operation. The other consequences of tool wear like increase in surface roughness, increase in temperature, increased vibration, decrease in the production efficiency, decrease in component quality, loss of tool, increase in cutting forces, etc could be identified only after some time.

Tool wear analysis research establishes the relationship with the machining parameters. Further, prediction and detection of tool failure based on the wear measurement is beneficial to avoid material losses and tool. A key prerequisite to the development of systems of such a capacity is reliable definition of the tool failure. One obvious criterion is that of the gross tool breakage. A better criterion than breakage for tool failure is based on the measurement of the tool wear and checking the tool condition. Wear measurement [2] is by far more popular, although other quantities such as tool surface finish have also been found to be sensitive measures of the tool life. The most commonly studied type wear is the flank wear, which is the scar that forms at the tip of the tool and easy to measure. The other types of tool wear are difficult to measure and are more
often encountered on the steel metal cutting tools than the carbide based tools.

Each type of tool wear tends to progress at varying rate. An initial period of the rapid wear, known as wear-in (stage-1), is followed by an extended period of the stable wear (stage-2). As the tool nears the end of its useful life, wear increases rapidly (stage-3) until the catastrophic failure occurs. In the stage-1, the wear occurs or grows rapidly with in the short period of time because during the initial contact of the sharp cutting edge with the work piece, the peaks of the micro unevenness at the cutting edge are rapidly broken away. In the stage-2 the wear progresses at the uniform rate. In the stage-3, the wear rate [4] is rapid and may lead to catastrophic failure of the tool. In general the most economical wear at which to remove the tool and re-sharpen is just before the start of the rapidly increasing portion of the curve, i.e. at the start of the region-3. The wear land on the flank will not generally be uniform along the entire cutting edge length. Although the flank wear measurements are popular for determining the tool condition, they are somewhat ambiguous. When tool wear is a complex geometric and metallurgical process, which can not be completely characterized by any signal numerical measure. Still, in the interest of obtaining meaningful results in a timely fashion, wear measurements can be considered useful points of reference to judge potential tool life estimation as a part of tool monitoring techniques.

6 APPLICATION OF NEURAL NETWORKS

Artificial Intelligence techniques are being used for tool wear analysis [13] Application of neural network by training of the network using suitable data is the basis for obtaining the conformity of analysis. Acoustic emission data such as ring down count, rise time, energy, event duration, peak amplitude were obtained from the experiments conducted and was used as input patterns. Further cutting forces were collected and were also used as input patterns. Data for each trial were also used as input patterns. The model of MLFFNN [7] used in the analysis which consists of the final storage after the network was trained. The training of the neural network was conducted for training space generated from the experiments conducted for Ceramics inserts on EN-31. Each training space consisted of 26 of seven parameters each thus forming a [26 x 7] input matrix. The training was conducted varying the momentum constant and hidden neurons, fixing the minimum error at 0.001mm to 0.003. Once the target was reached the weight were save and these weights can be used for testing for any given set of parameters. The network calculated the flank wear and also gives the conditions of the tool. These conditions are pre defined in the testing program. Depending on the flank wear the program interprets the tool condition and report as one of the cases for checking the condition of the tool namely Good, Average, Alarming State and Replace. This is a [7x26] training space pattern generated for ceramic at .09 mm/sec feed, .5 mm depth of cut and 500 rpm i.e., 7 parameters or input neurons and 26 input patterns.
6.1 Sample Training Space for Ceramic Tool at 0.09 Feed

<table>
<thead>
<tr>
<th>Function [ p, t ] = tti1data ();</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01[5474 3132 41420 32076 179536 12 07];</td>
</tr>
<tr>
<td>p02[3482 25751 37980 30082 119360 14 09];</td>
</tr>
<tr>
<td>p03[2963 24520 38442 30917 101481 16 10];</td>
</tr>
<tr>
<td>p04[3904 43067 54756 46000 134129 19 10];</td>
</tr>
<tr>
<td>p05[3498 27134 46933 39404 117309 14 07];</td>
</tr>
<tr>
<td>p06[2050 19789 43843 37907 066082 14 08];</td>
</tr>
<tr>
<td>p07[1659 10454 29999 25700 054633 15 14];</td>
</tr>
<tr>
<td>p08[1723 11358 34859 31458 062589 15 07];</td>
</tr>
<tr>
<td>p09[1832 13931 42513 60849 035659 09 04];</td>
</tr>
<tr>
<td>p10[4151 18633 36758 28680 013693 15 09];</td>
</tr>
<tr>
<td>p11[3626 18506 37732 30178 116019 16 09];</td>
</tr>
<tr>
<td>p12[3134 21086 36681 29023 105944 18 10];</td>
</tr>
<tr>
<td>p13[4365 24725 33565 28103 130532 16 10];</td>
</tr>
<tr>
<td>p14[1954 10076 40903 33131 064247 16 10];</td>
</tr>
<tr>
<td>p15[3194 25094 35453 27854 120769 19 11];</td>
</tr>
<tr>
<td>p16[2109 10013 35208 27535 065369 15 09];</td>
</tr>
<tr>
<td>p17[2102 11094 33589 27547 728095 15 06];</td>
</tr>
<tr>
<td>p18[2637 13179 34955 27574 085757 19 10];</td>
</tr>
<tr>
<td>p19[1579 13200 32595 27539 049048 20 07];</td>
</tr>
<tr>
<td>p20[1953 09588 28561 23683 060408 20 10];</td>
</tr>
<tr>
<td>p21[0240 01068 01859 01459 007934 20 09];</td>
</tr>
<tr>
<td>p22[4182 26107 29354 22906 127131 20 10];</td>
</tr>
<tr>
<td>p23[1954 11021 33622 27927 063410 18 10];</td>
</tr>
<tr>
<td>p24[3621 15069 35158 28199 118485 12 07];</td>
</tr>
<tr>
<td>p25[8557 16905 74033 56239 245899 16 09];</td>
</tr>
<tr>
<td>p26[3199 02859 31316 23931 103934 14 07];</td>
</tr>
<tr>
<td>t = [0.09 0.12 0.14 0.15 0.15 0.15 0.17 0.17 0.17 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2];</td>
</tr>
</tbody>
</table>

This is a [7x26] training space pattern generated for a ceramic tool at .09 mm/sec feed, .5 mm depth of cut and 500 rpm i.e., 7 parameters or input neurons and 26 input patterns.

7 DATA and ANALYSIS

Experiments were conducted to record the data. AE data recorded include Ring Down Counts (RDC), Rise Time (RTT), Event Duration (EDT), Energy (ENT) and Peak Amplitude (PA) for all the trials which also recorded Flank Wear for each trial along with the cutting forces measured with the help of lathe tool Dynamometer. Each trial also recorded any changes of the tool surface and material surface finish along with the measured parameter of Flank Wear.

7.1 A Sample data used for flank wear analysis

TOOL NAME: TTI15 (Ceramic) MATERIAL: EN31, FEED: 0.09 mm
DEPTH OF CUT: 0.5 mm SPEED: 500 rpm
7.2 A Sample data used for Acoustic Emission analysis

<table>
<thead>
<tr>
<th>Trial no</th>
<th>Forces in Kg</th>
<th>Flank wear (mm)</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X 12</td>
<td>Z 7</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>9</td>
<td>0.12</td>
</tr>
</tbody>
</table>

8 RESULTS AND DISCUSSION

Flank wear occurs primarily as a result of rubbing action between flank face of the tool and work piece surface. Specially, flank wear is chosen to study because of the fact that flank is directly involved in metal contact and is more prone to wear. It is strongly dependent on the speed, fed depth of cut, etc. in this experimental investigation, throughout; speed and depth of cut were maintained constant. The variety of AE signal produced is constant upon temperature and the extent of deformation in the zone. Therefore, AE signals are collectively influenced by the extent of tool wear, either directly or indirectly. The use of AE signal for general purpose monitoring of machining process is limited because the mechanism of generation of AE signal is not completely understood, and the analytical techniques for AE signals are still being evaluated.

Mere few trials may not be sufficient for a complete analysis. Also a short duration for machining operation does not suffice as AE parameter usually show higher value in the beginning when the tool comes in contact with the work piece and then stabilize during the cutting operation. The force significantly varies with respect to flank wear. Machining operation is speed dependent and tool wear is influenced by the cutting forces. As the cutting cycle is repeated, higher values of the cutting forces are recorded showing steady and progressive change in them.

RDC and Energy as seen in the comparison of graphs, which were generated for various tool inserts have close relation with the flank wear, at discerning points it can be seen that all the three parameters rise and drop simultaneously. Parameters of an acoustic signal which were considered for analysis i.e., energy and RDC were more sensitive to the progress of wear. Experimental investigation was carried out to assist the feasibility of AE signals for measuring and analyzing tool wear and estimate wear as a part of tool monitoring technique in the process of turning.

Acoustic emission signals are very sensitive to the sequence of signals during tool contacts, chip formation and tool engagement/disengagement leads to entry and exit peaks in AE signal during its cutting cycle. The change in the AE signal is due to more deformation and rubbing of tool materials. The AE signal clearly demarcates the cutting action between sharp edge and worn out tools. In the present study, it is observed that beyond 0.2mm flank wear; there is a rapid increase in AE parameters. Among the AE parameters RDC and Energy are more sensitive to the conditions of cutting tool.
Graph 1 and 2 shows the comparison of ring down counts of various materials at various feeds (i.e. 0.09 mm/rev, and 0.07 mm/rev). It can be seen from the graphs that RDC values are higher for higher feed rates which is because of higher wear of the tool tip.

Graph 3 shows the Comparison of Flank wear vs. Trial no. for various tools for one of the feed rates a feed rate of. It can be seen from the graph that the wear is higher at higher feed rates. Therefore it can be inferred that as the feed rate increases the wear rate also increases or in simple words the tip wears out quickly.

8.1 Comparative RDC graphs for different tool materials at same feeds

Graph 1: Comparative Cumulative RDC vs. trial no at .09 feed

Graph 2: Comparative Cumulative RDC vs. trial no at .07 feed
8.2 Training the Network for Ceramic Tool at .09 mm/s feed

The initial result of training the network could be obtained by using the error versus the number of epochs, which illustrated that the graph moves closer to X-axis as the number of epochs increases thus suggesting that probability of reaching the performance goal increases with increasing number of epochs.

Number of neurons in Hidden layer = 40
Momentum constant = 0.85
Maximum epochs = 10000
Error goal = 0.001

The performance goal was reached after 4616 epochs. The network thus trained and the final weights were saved for future testing.

8.3 Flank wear analysis of Ceramic Tool at .09mm/s feed

For a particular test it was found that the tested parameters tallied with that of the results obtained by training of neural network using AE signals. Actual flank wear recorded by direct measurement was 0.23mm, result obtained by training of Neural network was 0.2299mm and physical examination and result obtained based on the attributes of the trained network indicated the condition of the tool as Average. The flank wear obtained from the experiment measured directly using tool makers microscope were correlated with computed values of flank wear for 19th trial here. The results tally here and for all cases that were examined. With a suitable provision for online measurement of flank wear, Neural Network model may successfully be implemented in Computer Integrated Manufacturing Systems.

9 CONCLUSION

Experimental results obtained on measurement and the trained network generated attributes correlate for one complete experiment. Higher degree of accuracy may be achieved when the
training of network is taken up with more number of trials and use of larger number of epochs. Cutting tool wear proceeds the same way as the standard wear pattern indicating the right direction of the experiment. More than one method [8] of wear analysis and correlatives studies is being considered for obtaining higher levels of accuracy and for optimization.

10 REFERENCES

Multi-model Multi-strategy Teaching/Learning in Science, Engineering and Technology

Pradip Peter Dey1, Mohammad Amin1, Michelle Bright1, Arun Datta1,2, Shakil Akhtar3, Albert Cruz1, Mudasser F. Wyne4, Patrick Kennedy1, Patrick Olson1, Rell Snyder1, Arirenza Farahani1, Samuel Afuwape1, Hassan Badkoobehi1 and Amber Lo1

1National University, 3678 Aero Court, San Diego, CA 92123, U.S.A.
2 National University Community Research Institute, 11255 North Torrey Pines, La Jolla, CA 92037
3Clayton State University, Morrow, Georgia, U.S.A.
4University of Michigan, Flint, Michigan, USA,
pdey@nu.edu; mamin@nu.edu; mbright@nu.edu; adatta@nu.edu; sakhtar@clayton.edu; acruz@nu.edu; mfwyne@umflint.edu; pkennedy@nu.edu; polson@nu.edu; rsnyder@nu.edu; afarahan@nu.edu; safuwape@nu.edu; hbadkoob@nu.edu; alo@nu.edu

Abstract: Critical teaching/learning issues are raised considering all instructional strategies including lectures, transformations, experimentations, problem solving, analogical, case-based and mathematical reasoning. Recent trends and best practices are considered.

1 INTRODUCTION
Students as well as educators enjoy Problem Based Learning (PBL). With PBL, learners are usually organized into groups, one or more problems are given to each group for solving the problems under the supervision of the instructors and learning is driven by the problems [1, 2, 3, 4, 5]. Do we need to consider any alternatives given that PBL is so successful in science, technology and engineering? Is there anything to gain from combining multiple learning strategies? Are multi-model multi-strategy learning approaches feasible in science, engineering and technology?

2 CHALLENGES
Although PBL is highly successful in certain environments, it is not appropriate for all learners for all topics, since teaching methods are not dynamically adjusted on the basis of different challenges faced by different learners for different topics. Adjusting teaching methods based on learner feedbacks may be appropriate in multi-model multi-strategy learning environments. Agility in teaching/learning and grading helps to overcome the different challenges faced by different learners for different topics. According to Glickman "Effective teaching is not a set of generic practices, but instead is a set of context-driven decisions about teaching. Effective teachers do not use the same set of practices for every lesson . . . Instead, what effective teachers do is constantly reflect about their work, observe whether students are learning or not, and, then adjust their practice accordingly [6]. Our ability to teach is enhanced with our agility to adjust. In education, the same size does not fit all. In technologically advanced societies, problem based learning is usually practiced in combination with technology based learning [7, 8] at least in the medical domain because changes in medical technology were significant during the past 80 years that it is impossible to imagine solving problems in the medical domain without technology. Technology enhanced learning is receiving reasonable considerations around the world [9]. In any domain, in addition to PBL and technology based learning, one may consider any combination of game-based learning [10, 11, 12], community-based learning [13, 14, 15], work-based learning [16, 17], inquiry-based learning [18, 19], project based learning [20, 21, 22], brain-based learning [23, 24, 25], team-based learning [26, 27], internet-based learning
Understanding of the mechanisms underlying human information processing system is crucial for comprehension of learning. Mathematical reasoning is used in science, engineering and technology with advantages in multi-cultural environments with diverse learners. Christians, Muslims, Hindus, Jews and atheists may differ on many social or religious issues, but must agree with mathematical reasoning such as:

\[
\text{if } 3X = 12 \\
\text{then } X = 4
\]

which is crucial for science, engineering and technology. Humans use analogical reasoning in learning for certain topics in certain circumstances. However, its use may or may not bring disagreements among diverse learners or community of learners based on their religion, culture tradition or beliefs. Case-based reasoning and anecdotal reasoning are also used in learning.

3 IMPORTANT STEPS
The goals of the 21st-century academia are to provide the flexibility, improve the learning outcomes and equip learners with appropriate knowledge for solving more complex problems never experienced before. As we learn more about learning we understand scientific aspects of learning based on the recent contributions from neuroscience, psychology and cognitive science. It would be wonderful if knew more about learning. Can excellent teaching abilities be learned? Can these abilities be taught? What parameters characterize symbolic learning that can be isolated? Can we positively correlate and verify organic level learning? We are struggling with many such questions about teaching and learning. “Recent test results show that U.S. 10th-graders ranked just 17th in science among peers from 30 nations, while in math they placed in the bottom five. Research suggests that a good teacher is the single most important factor in boosting achievement, more important than class size, the dollars spent per student or the quality of textbooks and materials”[51]. Our discussion of learning considers all instructional strategies including lectures, transformations, experimentations, problem solving, analogical, case-based and mathematical reasoning for affective learning utilizing tools and technologies in an innovative way. We may combine several strategies in creative ways for the benefit of diverse learners. The practitioners realize that certain propositions are more important than others:

1. Learning outcomes are more important than tools.
2. Analyzing feedbacks from students is more important than declaring student centered environments.
3. Dynamically adjusting teaching strategies to learner’s goals and preferences is more important than following a teaching plan.
4. Teaching activities are driven by realistic problem solving, feedbacks from students and learning outcomes rather than by a schedule.
5. Demonstrating problem solving strategies on sample problems is more effective for teaching/learning than lecturing on them.
6. Dynamically combining multiple strategies in multiple models is more effective for teaching/learning than relying on a single pre-planned strategy.
7. Mathematical reasoning is more important than case-based reasoning.
One other aspect of education using multimodal techniques is how best to use technology for instructional purposes. There are several inherent constraints that educators must take into account when lecturing, presenting or designing learning environments. Three of which are listed in *Beyond Bullet Points* [57]:

1. Limits of working memory
2. Address two channels of learning
3. Attention guiding

This reference gives a model on how to construct a convincing and longer remembered presentation taking advantage of educational research in how humans learn. Knowing these constraints, how to avoid them and take advantage of how the human mind learns increases the fulfillment of the learning objectives in any modality.

One of the important challenges every teacher faces is how to combine multiple strategies in multiple models in order to serve 21st century learners in a reasonable way so that their learning goals are met within the scope of learning theories [58-68].

References:

[14] Community-Based Learning, [http://www.nwrel.org/scpd/sirs/10/t008.html](http://www.nwrel.org/scpd/sirs/10/t008.html), Retrieved February 10, 2008.
Exploration Strategy Based on Terrain Coverage for Multiple Robots Using Genetic Algorithm

K.S. Senthilkumar and K. K. Bharadwaj
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi 110067
ksskumar16@yahoo.com, kbharadw@gmail.com

Abstract

This paper considers motion planning approach based on terrain coverage for Multi Robot Exploration. The key problem to be solved in the context of multiple robots is to choose appropriate target points for the individual robots so that they simultaneously explore different regions of the environment. Here we are using Genetic Algorithm (GA) to assign the targets for the individual robots through which the coordination within multiple robots is achieved. A combination of cost and utility methods is used for fitness calculation. The cost means the cost of traveling the distance to the target and the utility of a cell means the size of the unexplored area that a robot can cover with its sensors upon reaching the target location or near by location. The terrain coverage technique is used to cover the visible environment by the robot. The proposed exploration strategy is implemented and tested on simulation. We demonstrate that our approach is well suited to control the motions of a team of robots in various environments by using different experiments

Keywords: Genetic Algorithm, Terrain Coverage, Multi Robots, Cell Decompositions, Motion Planning.

1. Introduction

Robots are widely used in domestic and commercial applications, typically involved with a specific task. Autonomous exploration and map-making become increasingly strong on single robots, but the next challenge is to extend these techniques to multiple robots. The coordinated multi robots exploration is a fundamental problem in robotics. The use of multiple robots is often suggested to have several advantages over single robot systems [3]. First, cooperating robots has the potential to accomplish a single task faster than a single robot [8]. Furthermore, using several robots introduces redundancy. Teams of robots therefore can be expected to be more fault-tolerant than only one robot. Merging of overlapping information, that helps compensate sensor uncertainty is another advantage of using robot teams [2]. However, when robots operate in teams there is a risk of possible interferences among them.

There are applications, which require a coverage path planning algorithm that specifically emphasize the space swept out by the robot. Integrating the robot’s footprint along the coverage path yields an area identical to that of the target region. This problem is related to the Covering Salesman Problem (CSP), where a robot must visit a neighborhood of each city that minimizes
the travel length [1]. One of the major activities of several recent works in coverage is the guarantee when the planner generates a path that completely covers the free space. Many terrain coverage planning algorithms use a cellular decomposition of the free space. A cellular decomposition means breaking down the target region into cells and complete coverage is achieved by visiting each cell in the decomposition. Terrain Coverage Planning (TCP) enables applications such as scouting and cleaning [12], robotic demining, snow removal, lawn mowing, etc.

This paper discusses about how coordination among the robots is achieved by assigning targets through GA technique. Moreover, for fitness calculation, it uses two parameters: the utility of unexplored areas and the cost for reaching these areas. This cost and utility method has been integrated already in [2]. But here it has been defined some specific points as targets in the occupancy map [14]. Many classical path planning algorithms and some coverage approaches assume that the robot knows the layout of the environment prior to the planning event. One benefit of map-based approaches is that a path can be generated more efficiently using a map. This paper considers TCP as an approach that determines a path for a robot to pass over all free points in the partially known plane space. The robot is capable of exploring the environment on its own.

The remainder of this paper is organized as follows. Section 2 introduces the background about the techniques, while section 3 discusses the problem specification and describes the cost & utility functions, target selection and cell exploration in details. Section 4 presents implementation and simulation results. Finally, section 5 discusses some conclusions and proposes possible future work.

2. Background

2.1 Genetic Algorithm
A new approach in robotics is evolutionary robotics, which uses evolution as a tool to create better robot controllers. Genetic Algorithms (GAs) [7] have shown to be powerful procedures for solving complex and multi criterion optimization problems. These algorithms mimic models of natural evolution, having all the benefits of a population based search method. Population represents a solution and each chromosome of the population is given a fitness value computed by an evolution function. When all the chromosomes in the population have been evaluated, reproduction, crossover, and mutation operators are applied to create a new population. In all generation, two best parents are selected, and one-point crossover is performed with a given probability. Then, values of the newly obtained strings are mutated with a given probability. Thus, in this evolutionary system, the new population replaces the old one maintaining only the best offsprings. After a number of generations, this process terminates when the best chromosome that represents the near-optimum solution is found.

There are many differences between GAs and traditional search algorithms; one, traditional search algorithms work with one solution at a time while GAs work with many solutions. Two, traditional algorithms are often deterministic while GAs are probabilistic in nature. GAs have some weak points as well. They are very slow when compared with other traditional algorithms, and also there is no guarantee that the solution is efficient enough. Considering the advantages
and weaknesses, one may conclude that GAs can be of greater value to those problems for which there is no optimized search algorithm. This paper presents a new approach to the Target Assignment Problem (TAP) that describes a GA controller system which evolves chromosome with target points. These points define the target location to each robot. Four robots have been used hence the size of chromosome is 4. The chromosome contains target as values, so, first gene is assign to robot #1, and second gene is assign to robot #2 and so on. The chromosome structure is given as follows. In each generation a fitness value is assigned to each of the $n$ chromosomes.

\[
\begin{array}{cccc}
1 & 3 & 5 & 7 \\
\end{array}
\]

Figure 1 : Chromosome Structure

2.2 Terrain Coverage
Terrain Coverage is a motion planning approach that determines a path for a robot to pass over all the free grid points in the terrain. The terrain or the target region will be broken down to cells and complete coverage is achieved by visiting each cell in the decomposition. Sometimes, this is called Grid Decomposition and is the set of non-intersecting regions called grid cell whose union fills the whole environment. Besides its theoretical interest, the mobile robot covering has several important applications. In the past years the coverage problem has received much attention and several solutions have been proposed in the literature including single and multi-robot solutions. An early reference to the coverage problem can be found in the work of Oommen et al. [10], Rao and Iyengar [11], and Choset [4, 5]. Application areas includes automatic floor cleaning in supermarkets [6] and railway stations [13], harvesting in large fields, field demining [9], and snow removing vehicles. This paper focuses on the coverage problem of coordinating a mobile robot’s footprint efficiently. Typically the robot can cover each cell by moving back and forth through adjacent cells.

3. Problem Specification

The goal of an exploration process is to cover the whole environment in a minimum amount of time. It is assumed that at every point in time both, the map of the area explored so far and the positions of the robots in this map are known. To represent the environment we use Occupancy Grid Maps (OGM) [14]. Each cell of an OGM contains a numerical value representing the probability that the corresponding area in the environment is covered by an obstacle.

Genetic Algorithm technique is used to coordinate the robots by determining appropriate target locations to each robot. By directing a robot to an unexplored target cell, one can expect that it gains information about the unexplored area when it arrives at its target location. In this approach several robots will not move to the same target location and robots do their exploration independently. Therefore the robots are simple with limited sensing and computational capabilities. This paper in addition considers the cost of reaching a target cell and the utility of that cell for calculation of the fitness. For each robot, the cost of a cell is proportional to the distance between the robot and that cell. The utility of a cell depends on the number of robots that cover the cell.

3.1 Cost Calculation
The cost of reaching the current target cells is determined by calculating the optimal path from the current position of the robot to all cells. This is based on a deterministic variant of value

165
iteration, a popular dynamic programming algorithm. As mentioned above, the cost for traveling
to a grid cell is proportional to its occupancy value. The cost calculation is done independently
for each robot. The steps used to calculate minimum-cost path is as follows.

Initialization Step:
\[ V_{x,y} = \begin{cases} 
0, & \text{if } (x,y) \text{ is Robot 's position} \\
\infty, & \text{otherwise} 
\end{cases} \]

Updating step:
For all grid cells \( C_{x,y} \) do
\[ V_{x,y} = \min(V_{x+\Delta x, y+\Delta y} + \sqrt{(\Delta x)^2 + (\Delta y)^2} \cdot p(C_{x+\Delta x, y+\Delta y})) \]
Where \( \Delta x, \Delta y \in \{-1, 0, 1\} \), \( p(C_{x+\Delta x, y+\Delta y}) \in [0, \text{occ max}] \).

We consider \( \text{occ max} \) as 2 grid cells positions, which is the maximum visibility range of the robot.
This technique updates the entire grid cell’s cost (V) by the value of their best neighbors, plus the
cost of moving to these neighbors. Cost is equivalent to the probability \( p(C_{x,y}) \) that a grid cell \( C_{x,y} \)
is occupied times the distance to the cell. The computation of V is done separately for each
robot.

3.2 Utility Calculation
This paper assumes that if there is a robot that moves to a particular target cell, the utility of that
cell can be expected to be lower for other robots. Since sensors of a robot cover the terrain
around a target cell as soon as the robot reaches there, the expected utility of those cells in the
locality of the robot’s target point is also reduced. Initially each cell has the same utility values.
Whenever a target point is selected for a robot, the utility of the adjacent cells \( (a) \) in distance \( d \) (2
grid cell) from the target cell \( (t) \) is reduced, according to the probability \( P_{\text{vis}}(a,t) \). The formula for
the Utility Calculation is as follows:
\[ U_t = U_t - P_{\text{vis}}(a,t) \]
We use a linear function to represent \( P_{\text{vis}}(a,t) \) which is given as:
\[ P_{\text{vis}}(a,t) = \begin{cases} 
1 - \frac{\|a-t\|}{\text{dis tan ce}}, & \text{if } \|a-t\| < \text{dis tan ce} \\
0, & \text{otherwise} 
\end{cases} \]
Where the distance means the maximum distance that the robot can sensor from the target
position.

3.3 Target Selection
The targets are assigned to each robot using GA with the cost and the utility which have been
considered separately for individual robots. To decide the assignment of target for robots, an
iterative approach using Gas is used. In each generation, initially the targets are randomly
assigned to all robots and the fitness for that particular chromosome is calculated by using the
following equation.
\[ \text{fitness} = \sum_{i=1}^{n} U_t^i - V_t^i \]
Where $V_i^t$ is the cost of traveling by robot $i$ to target $t$ and $U_i^t$ is the utility of target $t$ which is assigned to robot $i$. Number of robots are $n$. The GA terminates when the best fitness is achieved for a set of targets. The approach puts the constraint which assures that there is no duplication of target assignments.

This process is repeated until the whole environment is covered. The algorithm is as follows:

1. Calculate the cost ($V$) for reaching the entire cells in the environment by all robots separately.
2. Initialize the utility ($U$) for all the cells to 1.0.
3. While there is some more unexplored area do
   a. Determine the target for all robots by using GA
   b. Reduce the Utility of all target points and the adjacent points in the covered area from the targets of all the robots.

   \[ U = U - p_{vis}(a,t) \]
4. end while

### 3.4 Cell Exploration

In order to explore a cell, the robot must enter in to it. Once robot is in a cell, it knows which of the 4 neighboring cells are available, which has obstacles and boundary. Each cell in the robot’s covered area must be visited at least once. The robot moves forward using left-hand rule. For each cell that is visited the robot tries to move to the adjacent unexplored cell, following a left turn over a forward step over right turn. When ever all 4 adjacent cells are explored or blocked, robot will try to return along the path until it comes across an unexplored cell. This process will continue until all the cells in the covered area is explored. For example, Figures 2a, b, c show the way in which a robot explores the area.

**Figure 2a**

**Figure 2b**

**Figure 2c**

**Cell Exploration**

### 4. Implementation and Results

This approach has been implemented by using visual C++ language and performed simulation experiments in a Pentium M machine. We used 4 robots starting in different arbitrary locations in the environment with 12 targets. Target are predefined and decided in a way they do not appear in the current visibility range of the robots. The experiment is designed to demonstrate the capability of proposed approach to efficiently cover the environment with multiple robots. Figure 3 shows the map of the environment with targets. The size of the environment is $15 \times 20$ grids and is represented by occupancy grid map. The experiment was carried out in an environment without obstacles and with static obstacle (a wall) in two different environments. We found that this approach can efficiently guide the multiple robots to coordinately explore the environment. Figures 4 and 5 depict the map of the environment after some time of exploration and the final map respectively.
5. Conclusion

This paper presents a technique for multi robots to assign targets using genetic algorithm and covers an environment using heuristic coverage algorithm. Fitness is calculated by using cost of reaching the targets and utility of the targets, in which cost and utility computed simultaneously. This approach prevents robots to explore the place that are already explored by other robots. The free space is represented by a method called cellular decomposition. Since each cell is identical and the robot visits every cell, the resulting cell path completely covers the work-area grid in optimal time $O(N)$, where $N$ is the number of free grid cells. This approach has been implemented and tested on different simulated environments. Experiments show that this algorithm is capable of coordinating multiple robots during exploration in an efficient way.

Despite these encouraging results, there are several aspects which need to be improved. It is assumed that the environment is partly known with static obstacles. In many situations, this assumption may be unrealistic. Instead, the robot must use its on-board sensors to acquire information about the dynamic environments with moving obstacles and perform coverage online.

References


ABSTRACT
Knowledge management systems are aimed to provide knowledge-intensive organisations with tools, and methods to better manage their knowledge capital. A great success is gained in the course of managing explicit knowledge in the form documented knowledge fragments. But, greater part of organisations knowledge is realised in tacit form which is volatile and hardly captured in a formal way. Managing this type of knowledge still represents one of the major challenges in knowledge management research. This paper proposes a knowledge model that caters for capturing both tacit and explicit organisational knowledge in the software-engineering domain.

Keywords
Knowledge Management, software, tacit knowledge, knowledge reuse.

1. INTRODUCTION
As a result of the emerging knowledge-based economy (or K-Economy), knowledge has become the strategic resource of organizations to compete and survive. In line with this trend, the introduction of knowledge management systems (KMS) is receiving greater attention. KMS aims at providing an organization the ability to benefit from its past experience in order to respond more effectively in present situations. Organizational knowledge could be distinguished as formal (explicit) and informal (tacit) knowledge. According to Conklin [4], Formal knowledge is the stuff of books, manuals, documents, memos, white papers, plans and training courses, whereas informal knowledge is the knowledge that is created and used in the process of creating the formal results. It includes ideas, facts, assumptions, questions, guesses, stories, points of view, etc. Notice that these forms of tacit knowledge are personal and only reside in experts’ heads. In short, it constitutes what Koskinen [8] describes as the practical know-how.

Traditional KMS has mostly focused on the collection and dissemination of explicit (formal) knowledge, which is structured and stored in electronic knowledge repositories. But, there is a growing recognition that tacit knowledge constitutes an essential part of the organisational memory, and it has to be considered as well. According to Zack [15], explicating tacit knowledge so it can be efficiently and meaningfully shared and reapplied, especially outside the originating community, is one of the least understood aspects of knowledge management. Few attempts have been made to manage or at least partially capitalise tacit knowledge, but these efforts suffer limitations in managing it adequately. Capturing tacit knowledge is more complex than the management of explicit knowledge for multiple reasons. For example, tacit knowledge is highly personal and organisations have no control on its sharing and exploitation. Also there are many psychological as well as technical barriers that work against capturing the full richness of tacit knowledge.

The rest of the paper is organised as follows: first, the importance of capturing tacit knowledge is highlighted. Second, an overview of technical and non technical challenges of capturing tacit knowledge is discussed. Third, the conceptual framework and components of the proposed model is explained. The paper ends with a conclusion and suggestion for further work.

2. THE IMPORTANCE OF TACIT KNOWLEDGE
Experience is an indispensable resource for any organization to perform better in handling upcoming similar situations. Unfortunately organizations do not control all knowledge assets they have. Great portion of organizational knowledge is realized in tacit form which usually resides in workers’ heads as personal belief (know-how). Unfortunately when a human expert becomes unavailable either temporarily or permanently, then his/her expertise becomes unavailable too. In many situations knowledge workers who have been ousted for various reasons, are called back for consultation. This situation simply occurs because part of the organisation’s knowledge is taken away and it has to be retained. In a worst scenario, the knowledge expert might move to one of the business competitors, which means partially disarming knowledge of the former employer. This is why in organizations where workers turnover is high such as software engineering, the codification of tacit knowledge becomes an extremely important issue. The importance of managing tacit knowledge is well expressed in the following quote by Cummings [3], “An organization’s knowledge walks out of the door every night – and it might never come back”. Unfortunately, recent huge investment in IT is focused more on the management of explicit knowledge with less emphasis on the part of tacit knowledge management. Johannessen et al [6] asserts that companies easily lose their competitive edge if they emphasis investment in, and use of IT without taking tacit knowledge into consideration.
3. BARRIERS TO CAPITALIZING TACIT KNOWLEDGE

3.1 Non-Technical obstacles

Knowledge management is not a technology problem [14, 10]. Since it involves the externalisation of human expertise, human factors highly influence the success of any KMS. These non-technical factors are mostly related to firstly, how to characterise human knowledge, and secondly, how to promote knowledge workers to proactively contribute to externalise their tacit knowledge in order to build a group memory system. Usually there is a tendency that knowledge workers are very reluctant to share their personal knowledge with others. Some workers regard their professional knowledge as a source of job security and they are often reluctant to share it. Other organisational practices might also contribute to workers reluctance to share knowledge. According to Kokkoniemi [7], in organisations where staff can be rewarded for having particular expertise, this might lead to staff do not want to share what they know. In addition, psychological problems might also influence the position of knowledge workers participation in KM settings. For example, in meetings where the tacit knowledge usually produced and reproduced, some workers hesitate to share their knowledge and this would adversely influence the effectiveness of meetings as a result. Becker [1] justifies this attitude by, “reluctance of some members to speak, due to their shyness, low status, laziness or controversial ideas” [1].

All these problems have its roots in human psychological behaviour. Therefore, we believe that the externalisation of tacit knowledge has to be carried out as part of a defined knowledge lifecycle. It should not be left as merely voluntarily act. For example, in software engineering, deliberations held in different reviews of software projects lifecycle including the post-mortem reviews have to be formally captured. Unfortunately, the importance of the integration between software process models and knowledge lifecycle has probably not yet been sufficiently understood.

3.2 Technical Obstacles

In addition to non technical obstacles that contribute to difficulties in recognising and capturing tacit knowledge, tacit knowledge is usually less structured and it does not conform to a standard format. Therefore, it can be hardly coded and captured using traditional data models. From the software engineering viewpoint, design rationale models such as DRL [9], QOC [11] and IBIS [13] are all aimed to codify tacit knowledge in the form of different alternatives and justifications behind developers’ design decisions. In spite of the support of capturing tacit knowledge as facilitated by these models, they cannot accommodate the full richness of tacit knowledge. Sometimes, people are expected to articulate their tacit knowledge differently; they might even resort to the use of body language. There is no one model that can accommodate all these variations in expressing tacit knowledge. Another major limitation shared by all of these models is that, they are not coupled well with explicit knowledge in the form of knowledge embedded artefacts. We believe knowledge artefacts should be regarded as incubators of tacit knowledge.

The widespread use of the Internet made email systems an easy way to capture tacit knowledge. This approach can be exemplified by TeamInfo[2]. But, this approach still has limitations in the form of the lack of face to face interactions where richer tacit knowledge is likely to surface. Another major problem of email-based approaches lies in the fact that, the captured knowledge (in the form of email messages) is not sufficiently accessible by computational means because they do not conform to a standard knowledge model (i.e. free text). Therefore, we believe there is a need for knowledge models that provide decent characterisation of tacit knowledge and to integrate both tacit and explicit knowledge fragments. The target models also have to cater for managing the updottedness of captured knowledge based on a defined lifecycle. These requirements we believe will have potential to make our model workable.

4. THE PROPOSED KNOWLEDGE MODEL

A model is a purposeful abstraction of some part of reality [5]. It is meant to encapsulate the conceptual framework of the reality being modeled. Basically, our model is built around the notion of Knowledge Asset (K-Asset) as the basic building block of the potential software-engineering knowledge repository. K-Assets represent the smallest level of granularity of the knowledge repository. From the perspective of the software engineering domain, any useful fragment of software development knowledge can be regarded as a K-Asset. For example, any lesson learned or process model or knowledge-embedded software artefact can be considered as K-Asset. Individual knowledge workers can propose what they consider as valuable software engineering K-Assets. These K-Assets are then scrutinized by peer knowledge workers subsequently.

Figure 1 shows the higher-level abstraction of the proposed knowledge model. It shows the stages of the lifecycle of software engineering knowledge. Based on the proposed knowledge model, the knowledge generation loop starts from characterizing the building blocks of the experience knowledge. We believe the building blocks of the software experience are the knowledge workers and the set of software/hardware competencies they master and the knowledge-embedded software artefacts they produce. Any knowledge-embedded software artefact is in fact generated using one or a combination of software development skills in the form of software/hardware competencies. As a middleware between the knowledge workers and the knowledge-embedded software artefacts is the process models that needs to be followed. For example, in the case of any C programmer (i.e. knowledge worker), S/he must have mastery level of C programming (i.e. software competency). S/he can then follow the structured programming concept (i.e. process model) in order to produce any piece of software code. Supposing that the produced software code is an error trapping routine since this reusable code has the potential to be embedded in any C based software. This code can then be characterized as reusable knowledge fragment (i.e. knowledge-embedded software artifact).

Of course nothing can be characterized as a “knowledge fragment” simply based on the assessment of individual knowledge workers. Sometimes, the characterization of what can be considered as knowledge asset is a controversial issue because different knowledge workers might disagree about the validity of certain knowledge assets. What might be proposed by a knowledge worker as a knowledge fragments might not be qualified as such by others who might be more knowledgeable.
Moreover, the changing of circumstances and newly emerged facts might influence the context of previously characterized knowledge assets. This is the reality of the knowledge phenomena anyway. For example, at one point of time, Ptolemy’s theory (c. 2nd century CE) that we see objects because of the rays of light emanating from the eyes was very acceptable. With time and greater understanding, Ptolemy's theory was contradicted by ibn al-Haytham (c. 965-1040), who is regarded to be the father of modern optics. ibn al-Haytham correctly argued that humans see objects because the sun's rays of light are reflected from the objects into our eyes. This proves the fact that knowledge validity is dynamic. Therefore, based on the proposed model, the process of qualifying the proposed knowledge fragments relies on the consensus of other knowledge experts. This consensus is realized through other experts’ opinions (i.e. arguments) in relation to qualifying or disqualifying the proposed K-Assets. These arguments include qualifying/disqualifying criteria posted by peer knowledge workers make up the context associated with the captured knowledge. This collaborative knowledge filtering constitutes a pool for generating tacit knowledge in the form of views, suggestions, and all types of arguments (see the shaded portion in figure 1).

Figure 1. Conceptual view of the software-engineering Knowledge Model
Since, knowledge may become useless when the situation changes. Therefore, any captured knowledge has to be continuously examined and the relevance status of individual K-Assets changes accordingly. This goes inline with Nonaka’s claim that a company is not a machine but a living organisation [12]. We also believe that an organisation is living in the sense that it has to continuously examine what it knows and change its business beliefs accordingly. In the case of the proposed knowledge model, tacit knowledge generated through the collaborative knowledge filtering helps to maintain the updatedness of the captured knowledge. Figure 2 shows UML state chart diagram that demonstrates how the captured tacit maintains the dynamism of the relevance of captured K-Assets. Newly proposed K-Assets are considered inactive and only become relevant as it receives the agreement and support from other knowledge workers. Likewise, the more objections it receives the less relevant it becomes. The model discussed has been implemented through LiSER which is a software-engineering knowledge management system. It is meant to capture not only explicit knowledge but also tacit knowledge in the form of frequent arguments posted by knowledge workers (see Figure 3).

5. CONCLUSION
Knowledge is the winning factor for today’s organisations and knowledge management systems constitutes the tool through which organisations know what they know and what they should know. Unfortunately, knowledge management systems were structured more on formally structured knowledge while less attention is paid to capturing tacit knowledge. Software engineering is a very knowledge-intensive domain and a great portion of software organisations’ knowledge is realised in tacit form. It is usually held in professionals’ heads as practical know-how that might be lost at any time. Therefore, there is an urgent need to capitalise tacit knowledge as well. This paper proposes a knowledge model that caters for capturing both tacit and explicit knowledge in the software-engineering domain. The model integrates both explicit knowledge in the form of software artefacts and tacit knowledge in the form of arguments that constitute the context behind the creation and validation criteria of captured knowledge. In order to maintain the validity of captured knowledge, a knowledge lifecycle is defined where the captured knowledge is continuously examined by peer knowledge workers, and the relevance of respective knowledge assets are determined accordingly.

6. REFERENCES


---

**Figure 3. An implementation of the proposed model in the form software-engineering knowledge management system**

K-Asset main attributes

Explicit knowledge available through the URL provided.

URL of external explicit knowledge

Tacit knowledge in the form of argumentation map
On Simplification of Grid Scheduling and Resource Discovery Problem
A Survey and Proposal

Waqas Ahmed, Shaleeza Sohail and Aasia Khanum
Department of Computer Engineering
College of E & ME, NUST, Rawalpindi, Pakistan
waqas.jk@gmail.com, (shaleeza,aasia)@ceme.edu.pk

Abstract

Resource discovery and scheduling in grid environment is an extremely difficult task due to 1) inherent heterogenous nature of resources, 2) unavailability of single, central administrative domain, 3) unpredictability in joining and leaving of users, networks and resources, and 4) variability in availability of resources. In this paper we are presenting a survey of the existing resource discovery and scheduling approaches in grid. Secondly, we are proposing the idea of simplifying resource discovery and scheduling problem by using two phase approach. In first phase when large amount of resources are available, the approach uses fuzzy logic, which requires un concise information about resource availability for scheduling to reduce the resource discovery overhead. When the resource availability falls below a predefined threshold value, the approach goes into second phase and uses precise values for scheduling.

Keywords:- Resource Discovery, Scheduling, Fuzzy Logic, Grid

1 Introduction

Grid computing system provides a virtualized computing environment for dynamic selection, sharing and aggregation of multiple type of resources based on availability, capability, performance and cost [1]. Numerous types of components are part of grid like computers, storage, network, data, online instruments etc. These components share different resources to fulfill users’ requests like CPU cycles, files, bandwidth, memory etc. Some of the resources are highly dynamic in nature with respect to their availability like CPU and bandwidth. Lack of single administrative domain is an inherent characteristic of grid environment which introduces issues like resource heterogeneity, different ways in usage, scheduling policies, security mechanisms etc [2]. In addition, the users, networks and resources can join or leave the grid at any time. Due to these issues, resource discovery and job scheduling are the most important management tasks in grid environment.

We are proposing an idea of using two phase approach for resource discovery and scheduling. In first phase, when a large amount of resources are present, the resource discovery protocol only provides imprecise information about resources. The scheduling algorithm
uses this information for job scheduling using fuzzy logic and keeps track of the allocated resources. As the amount of available resources reaches a predefined threshold value, the approach transitions to second phase, where concise resource information is gathered by resource discovery protocol, which is used by the scheduler to allocate resources to the jobs. Moreover, in case a large amount of requests are received for same type of resource, the approach transitions to phase two and collects fine grained information about the resource availability before scheduling. In this way the overhead of resource discovery can be reduced considerably.

The anticipation of this approach comes from the fact that due to the dynamism in the resource availability in grid environment, no matter how precise the resource information is collected, the availability can change at the time of actual job deployment. A special emphasis is on the network resources on the Internet which cannot be reserved when the grid resources are connected via Internet, hence, the availability cannot be guaranteed. Depending upon the degree of dynamism of any resource, different threshold values are to be predefined. The approach transitions from one phase to another depending upon those threshold values.

## 2 Literature Survey

The amount of work done in the areas of resource discovery and scheduling in grid environment is huge. Here we are only discussing the works that are related to our approach in two relevant areas, resource discovery and job scheduling.

### 2.1 Resource Discovery

In [11], the authors present a resource allocation model based on agents, where different types of agents i.e., Job Agent, Broker Agent and Resource Monitoring Agents are used. Job Agents are mobile and Broker Agents are static entities which adds the benefits of both dynamic and static structure of Grid. This model uses market based strategy to allocate the resource to jobs. The proposed algorithm [11] is experimented for three topologies, tree, random, and single step. Results show that as the number of job Agents decrease or number of nodes increase, the performance starts to decrease. Rejection ratio of jobs also rises when number of jobs increases or requirements of resource increases. From the result we also came to know that this algorithm performs well in tree based topology as compared to random or single step.

In [12], the authors concentrate on defining the meta data for resource for multiple attributes complex query. According to authors, currently DHT and LDAP provide limited support for complex and multiple attribute queries for resource. RDF [13][14] is used with hierarchical structure and routing algorithms. To evaluate the performance of system three component of routing algorithm i.e.; Bloom Filter, Inter-cluster routing algorithm and intra-cluster routing algorithm are integrated and tested [12]. The routing algorithm is compared with two already existing algorithms Gnutella flooding and Random Walk. From the result it appears that the routing algorithm creates less query messages as compared to other two algorithms, hence, decreases query message overhead on the network.

The authors in [15] propose a design for Grid Information Service (GIS) that supports multiple attribute queries in distributed environment. The authors proposed algorithms can be used for load balancing on multiple information service servers [15]. To evaluate the algorithm two data sets are used. A synthetic data set [20] for experimental evaluation
and data set from Planetlab [21] for evaluation on real world data. The results show that the average number of hops for query is increasing slowly. Moreover size of the routing table also increases with the increase in size of network. Results also show that for load balancing, attribute splitting technique performs well against k-d tree. When number of attributes decrease average number of hop for query increases. With DHT a query with range attribute results in flooding the nodes.

In [16], the authors propose agent based resource management model. The model is based on three types of agent, Resource Provider, Resource consumer and Resource broker and uses market based strategy to allocate the resource. In the model authors address three concerns: agent oriented architecture, efficiency of resource negotiation protocol, and autonomic nature of grid resource management components.

The authors in [17] combine the peer to peer methods for resource discovery and management with grid strategies, which suggests a decentralized approach for resource discovery for large dynamic grid environments because hierarchical and centralized approaches are only good for static or less dynamic environments. The proposed algorithm is tested against different sizes of Grid to evaluate its effect of different parameters such as Time to Live (TTL), mean cluster size, response time, number of nodes in cluster etc. The experiments show that as cluster size increases, number of results from external queries also increase which attributed to the increase in processing load to find out these results. To establish a compromise between processing load and maximum number of results discovered, authors propose an index ratio (R). As mean cluster size increases, R also increases up to a threshold value after which its value decreases due to large number of local queries, addressed from within the cluster. The results show the average response time rises when value of TTL rises and decreases as size the of cluster increases [17].

In [18], the authors provide an insight into both peer to peer models and grid models. Authors propose that the basic difference between these two is their organization of resources. In grid, resources are organized but are changing while in peer-peer models resources are unstructured but are not changing much. Authors proposed the combination of the positive points of both models by using unstructured approach which can be applicable for dynamic resources while structured approach is applied for static resources [18]. The authors also propose a model extended from Chord protocol which adds additional parameter of content relevance. The new parameter is determined from the semantics of resource, defined along with resource. The authors discuss different semantic description defining languages such as XML, RDF, OWL, OWL-S and WSMO also for defining the resources.

In [19] authors propose Multiple Agent Grid Environment (MAGE) to cater for the problems associated with LDAP and DHT models. MAGE platform is divided into four fundamental components: agent management system, directory facilitator, message transport service and agents. Support for scheduling and monitoring is provided for local grid system which schedule and monitor local system and return result to main agent management system.

Some of the resource discovery strategies use P2P naming service which organizes the resources in the form of graph or hierarchy. This naming scheme finds the resource using its position in the hierarchy or graph. One of the major drawbacks of this scheme is that it does not support attribute-based resource identification [18] [3]. There are many strategies present with DHT although new modified DHT supports multiple attribute complex queries, but for large environments finding the resource for task may be time expensive.

To reduce the time for resource discovery we are proposing the use of fuzzy logic based
on existing profile of resources and pattern of requests to find out a set of resources that will match closely to the requirement. Further exact matching on this small subset will take much less time and also provides alternatives if one resource is not available or its status is changed during the discovery process.

2.2 Scheduling

In [9] authors focus only on the scheduling of resources for independent tasks (Meta-tasks) using fuzzy logic. In this approach, the task is split into independent subtasks and then weight of the resource is calculated using its attributes such as time, cost etc and finally allocates the task to resource using algorithm that uses fuzzy logic. The proposed algorithm depends on two parameters i.e.; task heterogeneity and machine heterogeneity. Results of the algorithm are compared with Min-Min, Max-Min, Opportunistic Load Balancing (OLB), and Fast Greedy approaches. The results show that schedule cost is better with FAA but the algorithm itself require more computing time, to calculate the schedule of tasks , than its counterparts.

In [7] the authors use fuzzy logic to schedule the task but concentrate on a more specific problem of uncertain completion time of tasks. The algorithm works well for the task which has predictable processing time. In the training mode, the algorithm uses sample data using crisp values and then it runs using fuzzy logic. Results show that Fuzzy logic approach performed well and scheduled the jobs better. This algorithm is tested for tolerance and the results show that it reduces the tardiness in scheduled task.

The authors of [22] focus on the global scheduling of tasks in multi clusters environment using fuzzy logic algorithm. The scheduling algorithm can also be used for parallel processing. The proposed approach is compared with Best-Fit algorithm using Matlab. The proposed algorithm uses job arrival time, number of tasks in a job, and execution to communication ratio as performance parameters. Due to the use of execution to communication ratio algorithm performs communication more efficiently and reduces the time as a whole for scheduling.

The authors [6] propose fuzzy logic algorithm applied to Globus Toolkit. Fuzzy logic algorithm uses CPU and memory utilization to schedule the tasks to required resources. Resource discovery algorithm and resource broker are also implemented for Globus Toolkit [6]. For evaluation the algorithm is implemented in Globus Toolkit 4 with four Grid systems. For experiment purpose Fibonacci sequence is computed on these grids using Fuzzy Logic Scheduler (FLS), Round Robin and also Random scheduling. The results show that FLS is better than its counterparts both in term of turnaround time and speed-up time.

The authors in [8] use particle swarm optimization algorithm to allocate resources to tasks. Particle swarm optimization technique is based on the position and velocity of particle for scheduling tasks. The proposed algorithm is compared with two other algorithms Genetic Algorithm (GA) and Simulated Annealing (SA). Results show that particle swarm algorithm computes and completes the job in less time as compared to other two algorithms.

The authors in [23] proposed an agent based approach for scheduling. For local grid load balancing instead of simple FCFS an iterative heuristic algorithm is used. For global load balancing a peer to peer service discovery and advertisement policy is used. Each agent’s coordination layer has knowledge of all the agent in the environment. This layer receives and store advertisements from other agent. To evaluate the performance of algorithm a number of parameters are defined for performance metric. These include total application execution time, average advance time of application execution completion, average resource utilization
rate and load balancing level. Experiments are implemented in PACE evaluation engine. Three experiments are conducted, first using FCFS policy, then using Agent based algorithm and last using resource discovery in addition to agent based algorithm. The results show that using the proposed algorithm the resource utilization is maximized as compared to FCFS. In addition, the application execution times are much better with proposed approach and as the number of resources increases the completion time also improves. From the result it appears that centralized approach is good for local grid load balancing.

The authors in [24] proposed an algorithm based on genetic algorithm which accounts for task’s constraints in the form of Direct Acyclic Graph (DAG). Genetic algorithm has the ability to come to a compromise in case of two conflicting resource providers. A new fitness function is defined as minimizing the turn around time and fulfilling the completion time of job.

3 Two Phase Approach

The main idea behind two phase approach is based on the observation that while a large amount of resources are available there is no need of collecting fine grained information about the availability of resources. However, less available resource requires constant monitoring before actual allocation can be performed. Hence, our approach collects imprecise or non-crisp values about resource availability and uses fuzzy approach to schedule and allocate resources at the time of high resource availability. In second phase when resource availability is low crisp information about resources is collected and jobs are scheduled based on those fine grained values. The approach transitions from one phase to another based on the threshold values of the resources, as shown in figure 1. Finding the appropriate threshold value for different resources is a complicated non-trivial task which considers multiple factors, for example:

- Type of resource
- Dynamism in availability
- Frequency of requests for that resource
In the following sections, firstly, a brief description of Fuzzy Logic is given. Secondly, the first phase of our approach is discussed in detail. During the second phase, we are planning to use an existing approach (discussed in section 2) depending upon the simulation study results that are part of our future work.

### 3.1 Fuzzy Logic

Fuzzy logic is a computational intelligence based generalization of the classical two valued logic, extending the classical definition of of set membership by allowing graded membership of set elements. A fuzzy set is created by assigning different membership grades to the various objects of the universe of discourse, depending on the degree to which each object belongs to the fuzzy set. Fuzzy sets provide an intuitive way to group several objects of the universe of discourse into a single semantic category along with an indication of how well a particular object in this group represents the concept symbolized by that fuzzy set. An innovative concept provided by fuzzy logic is that of a linguistic variable which is a variable that can take on linguistic or qualitative values. An example of a linguistic variable is Bandwidth which can take on values like Low, Medium, and High etc. Figure 2 illustrates the concept. Linguistic variables are a great convenience in systems where exact data is either not available or is not feasible to use because of various issues related to computational complexity. Linguistic variables are also used in approximate reasoning where they are employed by fuzzy IF-THEN rules for expressing the decision making process of the domain expert. These rules generally take the form:

$$IF \ X_1 \ is \ A_1 \ AND \ldots \ AND \ X_n \ is \ A_n \ THEN \ Y \ is \ B_n$$

where $X_i, i = 1, \ldots, n$ are the linguistic variables corresponding to the input domain, and $A_i, i = 1, \ldots, n$ are the allowable linguistic values for each of the $n$ input linguistic variables. Similarly $Y$ is the output linguistic variable and $B_i, i = 1, \ldots, n$ are its allowable values. The process of inference with fuzzy logic is shown in Figure 3.
As Figure 3 shows, the input to the fuzzy rule based system is a set of crisp values which are fuzzified by finding their degree of membership in appropriate fuzzy sets. These membership degrees indicate how much of the antecedent part of each rule is satisfied by each input variable, implying the degree of firing of each rule. The degrees of firing of various rules in the rule base are aggregated by using the fuzzy aggregation operators to get the fuzzy output of the system. Standard defuzzification operators are available to translate the fuzzy output to a single crisp value.

Our proposed approach uses fuzzy logic as an efficient paradigm to handle large amounts of QoS related data. Fuzzy sets offer a many-to-one mapping between several numerical parameters and their fuzzy counterparts. The use of fuzzy logic also helps the process of data representation; rather than maintaining large amounts of data corresponding to various application requirements and network attributes, a few fuzzy terms are sufficient to represent the entire domain. Moreover, the decision making and inference process is also simplified in the form of intuitive IF-THEN rules instead of elaborate mathematical models of the domain.

The use of a computational intelligence approach like fuzzy logic in network management is also made attractive by the fact that such a setup can always be customized by learning from examples and experience. In combination with other computational intelligence paradigms like neural networks and genetic algorithms, a fuzzy rule based system can be tuned to the available data. In this regard, we are exploring the use of Adaptive Neuro Fuzzy Inference System (ANFIS) for inculcating adaptability into the network management framework.

### 3.2 System Architecture for Phase 1

Linguistic fuzzy variables are used to denote the output and input variables of the system. The following fuzzy features, all normalized to the range [0,1] are used by the system:

#### Availability
This variable indicates the level of availability of a specific resource, these values are provided by the resource discovery protocol used in this phase. The following terms are used by this fuzzy set:

\[ T_{Availability} = \{Low, Medium, High\} \]

#### Requirement
The requirement of a certain resource by a specific request is expressed by means of this fuzzy variable. This variable has the following terms in its term set:

\[ T_{Requirement} = \{Low, Medium, High\} \]

#### Priority
The relative importance of a specific request with respect to other competing requests at that time, is expressed by this fuzzy variable. We are only considering the batch of request requiring resources at that time for scheduling. Any ongoing request will not be terminated or suspended, if resources are not available for the new request. Hence, any scheduled request has the highest priority. This variable has the following terms in its term set:

\[ T_{Priority} = \{Low, Medium, High\} \]

#### Independence
This variable indicates the degree to which a particular resource is independent of the availability of other resources while contributing to the successful completion
of the current request. This parameter is introduced to cater for the requests requiring multiple resources. We are assuming that the success of such requests depends upon the availability of all required resources and if any one is not available then the request cannot be fulfilled. This parameter has the following terms in its term set:

\[ T_{\text{Independence}} = \{ \text{Low, Medium, High} \} \]

**Optimality** This output variable indicates the suitability of allocating a particular resource for satisfying the requirements of a particular request. It has the following terms in its term set:

\[ T_{\text{Optimality}} = \{ \text{Low, Medium, High} \} \]

The above fuzzy variables are used in a fuzzy rule based system (FRBS) to assess the present performance of the network. For each \((\text{resource}, \text{application})\) pair, this FRBS uses rules of the form:

\[
\text{IF \ Availability \ is } \alpha \ \text{AND \ Requirement \ is } \beta \ \text{AND \ Priority \ is } \chi \ \text{AND \ Independence \ is } \gamma \ \text{THEN \ Optimality \ is } \delta
\]

where \(\alpha \in T_{\text{Availability}}, \beta \in T_{\text{Requirement}}, \chi \in T_{\text{Priority}}, \gamma \in T_{\text{Independence}}\) and \(\delta \in T_{\text{Optimality}}\)

The system output indicates whether the assignment of the current resource to the current request in the \((\text{resource}, \text{request})\) pair is optimal or not by using the linguistic terms in the term set of the output fuzzy variable \(\text{Optimality}\). Fig 4 shows the structure of the FRBS.

Fig. 5 shows the rule surface taking into account the effect of \(\text{Requirement}\) and \(\text{Availability}\) on the \(\text{Optimality}\) output of the system.
4 Future Work and Conclusion

For the evaluation of our approach we are planning to use the existing discovery and scheduling algorithms (some discussed in section 2) in two phases of the approach using grid extension for Network Simulator ns-2 [10]. The comparison of our two phase approach will be made to some of the existing approaches when used traditionally i.e; irrespective of the amount of available resources.

The paper provides a survey of some of the existing resource discovery and scheduling approaches for grid environment. A small description of the research work and discussion of the results are presented in the paper. It is evident from the literature survey that the resource discovery and scheduling is a complicated task which we are proposing to simplify using a two phase approach. In the first phase using fuzzy logic coarse grained scheduling is performed as resources are available in abundance. As the resource availability decreases, the second phase of the approach starts and performs fine grained scheduling.

References


[21] Ganglia Archives for PlanetLab at http://planetlab.millennium.berkeley.edu/


Joint Modeling and Optimization of Business Processes and Information Technologies

Igor Safonov and Vadim Safonov
CETERMCO, Arlington, VA inusin_1@msn.com

Abstract

The paper starts from the analysis of problems and risks in attempts to bridge the gap between business processes and automating them information technologies. In doing so, we tend to take into account achievements and troubles in management practice, automation experience, software engineering and marketing methodologies, as well as to take into consideration opinions of our opponents. Then, we propose successfully tested by our customers the most common approach to forecasting, planning and governing of engineering, management and reengineering for complex automated technologies, representing the symbiosis of business processes and information technologies.

Keywords: business processes, information technologies, modeling, optimization, trust engineering, and risk management.

The contemporary impact of Information Technology (IT) on Business Processes (BP) is in the state of confusion between unsatisfactory grounded methodologies and insufficiently efficient tools. The actual slogan of the Third Gartner Business Process Management Summit (San Diego, CA, 26-28.02.2007) “From Process Understanding to Business Innovation” speaks for itself. Corporate interests of information industry still prevail over customer satisfactory, and a quantity conquers a quality. Developers and managers of IT badly know demands of BP to automation, and businessmen along with other stakeholders not always understand potential and restrictions of business computerization. But the tendencies are encouraging: process-oriented horse take her stand as a wheeler between tired object- and data-oriented out-runners. Our many-year adherence to a process-oriented approach bears fruit and it makes us (and our respected customers – no complains) happy. In our metamethodology developed by the International Unity Science Institute (IUSI), domain experts use their knowledge and experience in BP metamodel decomposition, and software architects, designers and developers use their knowledge and experience in IT aggregation interfacing BP and IT “just-in-place” of “docking”.

Actual and potential customers of IUSI and affiliated with it startup the Corporate and Enterprise Trust Engineering & Risk Management Company (CETERMCO) frequently ask us about the most fundamental problem in process of bridging the gap between different goals of people creating and operating BP and people developing and implementing IT. Both of them need a
common unified language for description, modeling and optimization of any automated BP – simple and precise language understandable by all stakeholders of a business automation processes. The language must to have an algorithmic (procedural, behavioral) syntax invariant for creators, operators and owners of BP and engineers, managers and operators of IT [8].

The only one difference between BP and IT models consists in different interpretation of the language operators on the last steps of their details. More exactly, there are two (BP and IT) algorithm descriptions of business and data processing with the same syntax and semantics, but syntax, semantics and pragmatics of their operators principally different (business and information activities). We have such metalanguage (XAAL) and dozens years of successful experience of its application for optimization engineering and optimal management of MIS, CAD, CAM and DSS for businesses, industry, governments and society.

Why we concentrate our attention simultaneously on engineering and management when these problems are so different? Because an effective management can’t be realized without features embedded into business process after previous forecasting, planning, design, development, testing, verification and validation, and vice versa engineering must take into account resources of management. Shortly, engineering and management are interdependent parts of the same BP life cycle. The problem is how to optimally distribute available resources between engineering, management, maintenance and reengineering stages of the life cycle. Extending the gold words of Viktor Glushkov about disorder automation, we can tell that bad automation very seldom can be compensated by good maintenance.

Software methodologists, developers and vendors advertise different approaches to software engineering. We recommend for small number of data and simple data processing algorithms – the object-oriented approach, for very big volume of data and simple algorithms – data-oriented (database-centered) approach, for relatively small volume of data, but tremendously complex algorithms – (structured!) process-oriented approach. In more complex situations, you can combine different approaches for relatively independent subsystems. For future, we believe in success of the intentional [4], aspect-oriented [2, 6] and insertion [3] programming.

Selection of software tools for software engineering, management, maintenance and reengineering in big corporations is complex multi-criteria problem. Simple recommendations based on “As Is” and “To Be” analysis can’t help. For these cases, we created and effectively used the interactive expert-modeling methodology. Our services can be outsourced. Experienced customers can buy a license and in the most troubled situations may hope on our consultations. We have very diversified experience in using our methodology for products and services selections by industries and governments, for regional development assessment, for movie competition and price establishment, for educational and training goals.

The metamethodology is based on contemporary theory of comparative analysis and interactive algorithms of expert-modeling system. Our real and potential customers are industrial, commercial, financial, transportation, communications, healthcare, pharmaceutics and entertainment enterprises and corporations, government, regional and municipal enterprises and institutions. Of course, first of all we oriented on executive teams and individuals, but our
models and tools may be useful for business analysts, process architects and engineers, configuration and change managers; performance, quality, reliability, security, safety and trust professionals [1, 5, 9].

Our major methodological principle is to use optimal combinations of the best formal techniques with the best practical experience. For example, one of these techniques is the predecessor of all iterative methodologies – a round-trip process-oriented engineering with common for all stages of object (BP, IT, BP & IT) life cycle an optimization block based on unified algorithmic language and simple mathematical methods. The technique allowed us to create unified methodology of BP engineering, management, maintenance and reengineering that provides generation of optimization models from a BP and/or IT formal description (Analysis), execution of BP and/or IT optimization procedures (Optimization), and regeneration of optimized BP and/or IT formal description (Synthesis).

On every stage of a BP and/or IT life cycle, the methodology can be executed iteratively zero, one or more times. The number of iterations is optimized. Starting from a first time implementation of the automated BP, corresponding procedures become procedures of reengineering and they are incorporated into a BP management and maintenance. The iterative-optimization engineering of BP and/or IT was developed a lot of years ago and successfully used many dozens times in different automation projects. The technique was used as one of basic procedural components for the expert-modeling system SELENA.

The next principle is the strict compliance with one of the most important requirements of engineering and reengineering – a functional equivalency of initial and optimized BP. Our main framework was based on the third principle – the system approach to engineering and reengineering of automated BP and automating them IT realized in the multidimensional and heterogeneous space of Goals, Behaviors, Structures and Resources [7]. Although BP and IT have cardinally different Goals, Structures and Resources, but they implement functionally equivalent processes distinguished only by nature and temporary logics functioning of their components. The basic phenomena have been understood dozens years ago, and considerable time elapsed before they could be put to wide industrial and commercial use.

In 2007 alone, our metamethodology was used for 1) Software Tool Assessment and Modernization in Division of Information Technology of the major Insurance Corporation; 2) Creation of Supporting Techniques for Standardization of Desktops’ Configuration and their Capability Improvement for Government Departments and Agencies declared by the Federal Desktop Core Configuration (FDCC) program; and 3) Pilot Project Development for Testing and Evaluation of Virtual Framework and Process Approach for Software Evolution on base of a Federal Government Agency.

Everybody talks about Business Process Orientation, Modeling and Optimization. But the last two problems in majority of practical cases have never been sufficiently resolved. Many are of the opinion that a complexity of processes is an insurmountable barrier in the path of engineering and management. It is not universally true. Our knowledge and experience in the field give us
grounds to propose our metamethodology, models and services to any customers. You may be surprised but in case of more complex BP and IT our models work more effectively and optimization reaches appreciable results.

It is not a paradox but good known conformity to natural laws – the simplest problems of business, engineering and management activities are resolved better by experienced and talented people without any models and tools. But in case of thousands (sometimes millions and billions) potential situations of decision making and possible consequences of these decisions only mathematical models and based on them computerized tools may be trusted. Recent chess games between computers and world champions of last years also approved it.

Notorious visibility of flowcharts, UML, PERT, GERT and GANTT diagrams, Petri networks, etc. cannot intensify creativeness of executives, managers, engineers, programmers and any professionals higher than the restrictions of human nature allow. Many thousands years ago, people moved from pictures to words and numbers. Why we need to move backwards into the Stone Age?

References


Prediction Based Dynamic Load Balancing Techniques in Heterogeneous Clusters

Pushpendra Kumar Chandra, Bibhudatta Sahoo
National Institute of Technology Rourkela, Orissa
pushpendrachandra@gmail.com

Abstract—The goal of load balancing is to assign each node a number of tasks proportional to its performance. Many load balancers have been proposed that deal with applications with homogeneous tasks but applications with heterogeneous tasks have proven to be far more complex to handle. In this paper we present a load balancing techniques that can deal with applications with heterogeneous cluster. Here we are considering three type of load I/O, CPU, MEMORY. For efficient load balancing we are proposed a new techniques that reduce the average response time. For that we are using the prediction scheme that predict the requirement of task and then according to prediction scheme this techniques assign the task to best suitable node.

Keywords: Heterogeneous cluster, I/O intensive load, Load balancing

I. INTRODUCTION

Load balancing (LB) is a critical issue in parallel and distributed systems for the efficient utilization of the computational resources. There is a large body of literature on load balancing and all the proposed load balancing algorithms can be broadly characterized as static and dynamic. The focus of this paper is on the dynamic load balancing algorithms and the processing times of the jobs are known at the time of execution. Load balancing can be static or dynamic.

In static scheduling, the assignment of the tasks to the nodes is done before the execution of the program. Information regarding task execution time and processing resources is assumed to be known at compile time. A task is always executed on the node to which it is assigned.

Dynamic scheduling is based on the re-distribution of processes among the processors during execution time. Redistribution is performed by transferring tasks from heavily-loaded processors to lightly-loaded processors with an aim to minimize the processing time of the application. The advantage of dynamic load balancing over static scheduling is that the system need not be aware of runtime behavior of the application before execution. The flexibility inherent in dynamic load balancing allows for adaptation to unforeseen application requirements at run-time. The major disadvantage of dynamic load balancing schemes is the run-time overhead due to:

1. The load information transfer among processors,
2. The decision-making process for the selection of processes and processor for job transfers, and
3. The communication delay due to task relocation itself.

Dynamic LB algorithms can be further classified into a centralized approach and a decentralized approach. In the centralized approach only one node in the distributed system acts as the central controller. It has a global view of the load information in the system, and decides how to allocate jobs to each of the nodes. The rest of the nodes act as slaves; they only execute the jobs assigned by the controller. The centralized approach is more beneficial when the communication cost is less significant, e.g. in the shared-memory multi-processor environment.

The main motivation of our study is to propose a centralized dynamic load balancing algorithm that can cater for the following unique characteristics of practical distributed computing environment:

- Heterogeneous system: There may be a difference in the hardware architecture, operating systems, computing power and resource capacity among sites. In this study, heterogeneity only refers to the processing power of site.
- Effects from considerable communication delay: The communication overhead involved in capturing load information before making a dispatching decision can be a major issue negating the advantages of job migration. We should not ignore the considerable dynamic communication delay in disseminating load updates.

Most load balancers were designed to handle applications with homogeneous tasks, for example data parallel application or tree-based algorithms. A lot of applications however consist of heterogeneous tasks, i.e. tasks performing different operation or operating on different types of data. Due to uneven job arrival patterns and unequal computing capacities and capabilities, the computers in one node may be overloaded while others in a different node may be under-utilized. It is therefore desirable to dispatch jobs to idle or lightly loaded computers to achieve better resource utilization and reduce the average job response time.

The rest of the paper is organized as follows. In the section 2 that follows, related work in the literature is briefly reviewed. In section 3, we describe the system model. In section 4 we describe the novel load balancing algorithm. Finally concludes the paper by summarizing the main contribution of this paper.

II. RELATED WORK

There are many approaches to balancing load in disk I/O resource can be found in literature [1][2][3][4][6][10]. Xiao Qin [1] proposed an algorithm IOLB and compare this algorithm with conventional CPU- and memory-aware load balancing schemes and shows that the IOLB algorithm significantly improves the resource utilization of a cluster under I/O-intensive workload.

Mais Nijim Tao Xie, 2005 developed a performance model for self-manage computer systems under dynamic workload condition, where both CPU- and I/O-intensive applications are running in computer systems. They shows that the controller is capable of achieving high performance for computer systems under workloads exhibiting high variabilities.

Xiao Qin et al. [4] proposed a feedback control mechanism to improve the performance of a cluster by adaptively manipulating the I/O buffer sizes. The primary objective of this mechanism is to minimize the number of page faults for memory-intensive jobs while
improving the buffer utilization of I/O-intensive jobs. The feedback controller judiciously configures the weights to achieve an optimal performance. Meanwhile under a workload where the memory demand is high, the buffer sizes are decreased to allocate more memory for memory-intensive jobs, thereby leading to a low page-fault rate. Increasing attention has been drawn toward I/O-intensive application. Kandaswamy et al. [10] examined optimization techniques and architecture scalability. They evaluated the effect of the techniques using five I/O-intensive applications from both small and large applications domain.

Xiao Qin et al. [6] developed two effective I/O-aware load-balancing schemes, which make it possible to balance I/O load by assigning I/O-intensive sequential and parallel jobs to nodes with light I/O loads. However, the above techniques are insufficient for automatic computing platforms due to the lack of adaptability. We proposed a algorithm that take all the parallel task and it balance the I/O-intensive load with effective manner.

III. SYSTEM MODEL

In this study we have considered a cluster computing platform of heterogeneous system in which set of $N = \{N_1, N_2, N_3, \ldots, N_n\}$ n nodes are connected via a high speed network. Each node in this model composed of a combination of various resources including processor, memory, disk, network connectivity and every node is differ with their processor, memory and disk. A load manger or master node is responsible for load balancing and monitoring available resources of the node. Fig 1 shows the queuing model for load manager.

![Fig. 1. Queuing model for load manager](image)

Load manager or master node process all arrival task in a FCFS manner. Here we are assuming that all task are to load manager is poison process. After being handled by load manager task are dispatched to one of the best suited node for execution. The nodes, each of which contains a local queue, can execute task in parallel. Load manger is composed of three modules: (1) predictor; (2) selector; (3) scheduler;

When new task is arrives at load manger, the identification of program being executed is sent to the predictor which predicts the resource requirements of the task. These predicted values are then fed to the selector which selects the node with under utilized and well suited for its requirement resource.

Predictor is used to predict the file I/O, CPU and memory requirements of a task. For this we use prediction scheme described in [7] uses a statistical pattern-recognition to predict the task requirements. The prediction is made at the beginning of a process's life, given the identity of the program being executed.

The prediction scheme consists of two parts. In the first part, which is an off-line procedure, resource usage states are determined for program executions of a given UNIX system. Resource usage data is collected for all processes that ran on the system for a few days, this data is analyzed as follows: Each process is represented by a point in a three-dimensional space, where each dimension corresponds to the resources of the system, i.e., the CPU, the memory, and the file I/O. A statistical clustering algorithm is then used to identify the high density regions of this three-dimensional space (i.e., determine the number of such regions and the means of their centroids). By definition, most program executions occur in or near these regions, and therefore they are referred to as the resource usage states.

In the second part, which is an on-line procedure, actual prediction is made. The prediction scheme builds and maintains a state-transition model for each program on an on-going basis. The states of the model are the resource usage states defined above. Suppose a program has been executed several times, providing a sequence of execution instances. First, the sequence of execution instances is converted into a sequence of resource usage states by assigning the nearest resource usage state to each execution instance. The state transition probabilities are then calculated from this new sequence to build a state-transition model for the program. The prediction is a weighted mean calculation of resource requirements using the program’s current state-transition model and the actual resource usage in its most recent execution. See [7] for further details. Then predicted value is fed to the selector that is used to select the best node among all nodes where the task will execute. That node is under-loaded and gives response effectively. Scheduler is responsible to dispatch the task to the node selected by the selector. Then task will send to that node and task will execute there. Load manager update the load status table.

IV. LOAD BALANCING ALGORITHM

We proposed a algorithm for a wide variety of workload conditions including I/O-intensive, CPU-intensive and memory-intensive load. The objective of the proposed algorithm is to balance the load of three types of resources across all nodes in a cluster. In this study analytically evaluate the performance of algorithm; we are focused on a remote execution mechanism in which task can be running on a remote node where it started execution. Thus preemptive migrations of tasks are not supported in our algorithm.

To describe this algorithm first we introduce the following three load indices with respect to I/O, CPU, memory resources. (1) CPU load of a node is characterized by the length of CPU waiting queue, denoted as $L_{CPU}(i)$, to identify whether node i’s CPU is overloaded. (2) Memory load of a node is the sum of the memory space allocated to all the task running on that node. The memory load of node i is denoted as $L_{MEM}(i)$ (3) I/O load measures two types of I/O accesses, i.e. (a) implicit I/O request includes by page fault; (b) explicit I/O request issued from tasks. I/O load index of node i is denoted as $L_{IO}(i)$. TABLE I shows the definition of notation we used in this paper.

Now we describe the load balancing algorithm of which the pseudo code is shown in Fig.2. Given a set of independent tasks submitted to the load manager. Our algorithm make an effort to balance the load of the cluster resource’s by allocating each task to a node such that the expected response time is minimized.

For each task $t$ our algorithm repeatedly performs steps 2-19 described follows:
First it will predict all three IOREQ, CPUREQ, MEMREQ, requirements of task j from set of task by step 2. This three predicted
value are important because according to this value task execute with best suited node. Step 3 is used to find the highest requirements of task and it is responsible for initiating the process of balancing I/O resources. Step 4-7 are used to balance the I/O load. In step 6 If the I/O requirements of task j are high then it will find the set of node where I/O load is minimum and satisfy all three requirements. Step 5 calculates the response time of task with all selected node. Step 6 if the response time is minimum with particular node then task will send to that specific node.

**TABLE I**

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Number of node in heterogeneous system</td>
</tr>
<tr>
<td>t</td>
<td>Task submitted to the system</td>
</tr>
<tr>
<td>λ</td>
<td>Arrival rate of task</td>
</tr>
<tr>
<td>μ_n</td>
<td>Service rate of heterogeneous system</td>
</tr>
<tr>
<td>IOREQ_j</td>
<td>I/O requirement of task j</td>
</tr>
<tr>
<td>CPUREQ_j</td>
<td>CPU requirement of task j</td>
</tr>
<tr>
<td>MEMREQ_j</td>
<td>MEMORY requirement of task j</td>
</tr>
<tr>
<td>L^IO_a</td>
<td>I/O load on node(1≤a≤n)</td>
</tr>
<tr>
<td>L^CPU_a</td>
<td>CPU load on node(1≤a≤n)</td>
</tr>
<tr>
<td>L^MEM_a</td>
<td>MEMORY load on node(1≤a≤n)</td>
</tr>
<tr>
<td>L^k^IO</td>
<td>I/O load index on set of k node that satisfy all requirements</td>
</tr>
<tr>
<td>L^k^CPU</td>
<td>CPU load index on set of k node</td>
</tr>
<tr>
<td>L^k^MEM</td>
<td>MEMORY load index on set of k node</td>
</tr>
<tr>
<td>R^k_j</td>
<td>Response time of task on set of k nodes</td>
</tr>
</tbody>
</table>

Second, step 8 if the memory requirements of task are high then it will perform to step 9-12 to balance memory load among all nodes. Page fault behaviors occur when the memory space allocated by running tasks exceeds the amount of available memory. That’s why it is necessary to balance memory to minimize the page fault. Step 9 searches the set of node with minimum memory load and satisfies all three resource requirement of task. Step 10 calculate the response time of task with all selected node then step 11 find the minimum response time of task from selected node then step 12 dispatch the to selected node.

**Algorithm: Load balancing**

Input: a job with task j submitted to master node
1. for each task do
2. Predict the value of I/O, CPU and memory requirements
3. if \( \text{IOREQ}_j = \max(\text{IOREQ}_j, \text{CPUREQ}_j, \text{MEMREQ}_j) \)
4. choose set of k node such that node \( L^k_{aIO} = \min(\sum L^a_{IO}) \) satisfy the all three requirements
5. calculate response time \( R^k_j \) of task j in set of k node
6. if \( R^k_j = \min(R^k_j) \) then
7. dispatch the task to node Ni and execute there
8. else if \( \text{MEMREQ}_j = \max(\text{IOREQ}_j, \text{CPUREQ}_j, \text{MEMREQ}_j) \)
9. choose set of k node such that node \( L^k_{MEM} = \min(\sum L^a_{MEM}) \) satisfy the requirements
10. calculate response time \( R^k_j \) of task j in set of k node
11. if \( R^k_j = \min(R^k_j) \) then
12. dispatch the task to node Ni and execute there
13. else if \( \text{CPUREQ}_j = \max(\text{IOREQ}_j, \text{CPUREQ}_j, \text{MEMREQ}_j) \)
14. choose set of k node such that node \( L^k_{CPU} = \min(\sum L^a_{CPU}) \) satisfy the requirements
15. calculate response time \( R^k_j \) of task j in set of k node
16. if \( R^k_j = \min(R^k_j) \) then
17. dispatch the task to node Ni and execute there
18. update the load status;
19. end for

Third, step 13 is responsible if the CPU requirements of task is high and step 14 is search the set node with minimum CPU load among all node that satisfy all requirements of task. And then calculate the response time of task in each selected node. Step 16 find node that gives minimum response time to execute the task. Step 17 dispatches the task to the selected node. Last step 21 maintains updated load information that is send to the load manger

**CONCLUSION**

Even though there are number of different dynamic load balancing techniques for cluster systems, their efficiency depends topology of the communication network that connects nodes. This research has developed an efficient load balancing for I/O, CPU, MEMORY intensive for this we develop a new way to predict and calculate the load of cluster nodes. The proposed load balancing
scheme aim to achieve the effective usage of global disk resources in cluster. This can minimizes the average slow down of all parallel jobs running on a cluster and reduce the average response time of the jobs.

Future studies can be performed in following direction. First, we will evaluate the performance of scheme on a large scale of cluster. Second, we have assumed the task is independent, so we will also simulate this scheme for inter-dependent task. Third, in this study we have assumed network communication cost is negligible, therefore we will extend this to balance load in network resource.

ACKNOWLEDGMENT

This research was supported by R&D project grant 2005-2008 of MHRD Government of India with the title as “Fault Tolerant Real Time Dynamic Scheduling Algorithm For Heterogeneous Distributed System” and being carried out at department of Computer Science and Engineering, NIT Rourkela.

REFERENCES

[4] Xiao Qin, A feedback control mechanism for balancing I/O-intensive and memory-intensive applications on cluster, parallel and distributed computing practices journal
[5] Paul Werstein ,Hailing Situ and Zhiyi Huang , Load balancing in cluster computer, Proceeding of the seventh international conference on Parallel and Distributed Computing, Applications and Technology (PDCAT’06)
[8] Xiao Qin, An availability-aware task scheduling strategy for heterogeneous systems, IEEE transaction on computers
QUERY PROCESSING IN MAIN MEMORY OODB SYSTEMS

Mudasser F. Wyne
University of Michigan-Flint
Department of Computer Science, Engineering Science and Physics
Flint, MI-48502, USA
mfwyne@umflint.edu

ABSTRACT:

As a result of the wide acceptance of object-oriented database systems, performance related issues become a crucial factor in the success of object-oriented database systems. We proposed a main memory Object Oriented database (MOOD) system [4] in which the entire database is memory resident to meet this performance demand. In this system to improve performance the system architecture, data structures, and optimization techniques are designed to exploit the fact that database is memory resident. As database technology is applied to more and more application domains, user queries are becoming increasingly complex (e.g. involving a large number of joins and a complex query structure). Query optimizers in traditional disk resident database management systems (DBMS), are not developed for efficiently processing such complex queries as well as for system like MOOD, often suffer from problems such as long optimization time and poor optimization results. In this paper we propose a set of cost functions to measure query performances. Finally, as a case study, we apply these cost functions to retrieval operations involved in query processing for MOOD.

Keywords: Main Memory Database (MMDB), Object Oriented Database (OODB), Data Structure, Query Tree, Indexing.

1. INTRODUCTION

An object oriented database is a collection of objects whose behavior, state, and the relationship are defined in accordance with an object oriented data model. In this, a class is an object, and a class may have any number of subclasses. A class inherits attributes and methods from a superclass, thus classes form a hierarchy. Databases in general and object oriented in specific are getting more and more important for storing complex objects from scientific, engineering, or multimedia applications. Examples for such data are chemical compounds, CAD drawings, XML data, or color images. Object-oriented Database Systems (OODBS) which deal with complex objects require expensive traversal costs to process queries. Since the database field is a rapidly growing one, database technology is applied to more and more application domains. These wider range of applications make user queries more and more complex in terms of the number of operations (e.g. more than 50 joins) and the query structure (e.g. snowflake queries).

1 The work presented is supported by a research grant from the University of Michigan-Flint.
There are many studies on efficient query processing against nested objects in OODB Systems. To answer a query, involving class hierarchy, in the top-down approach the traversal between a predicate class (class involved in the predicate) and a target class (class from which required objects are retrieved) requires a high processing cost. To overcome this cost, several indexing schemes have been proposed, i.e. path index [1], tree [10] and path signatures and path dictionary.

One approach to improve performance of at least one type of queries namely, range queries is to use filter refinement architecture [8]. The core idea is to apply a filter criterion to the database objects in order to obtain a small set of candidate answers to a query. The final result is then retrieved from this candidate set through the use of the complex similarity measure. This reduces the number of expensive object distance calculations and speeds up the search process. Another possibility is the use of a metric index structure. In [9], several efficient access methods for similarity search in metric spaces are presented. In most real-world applications a static index structure is not acceptable, so dynamic index structures like the M-tree [10] are applied. Authors in [2, 8] claim that these concepts can beneficially be combined and that through the combination a significant speed-up compared to both separate approaches can be achieved.

In an effort to improve performance of query processing, query optimization is also a topic of interest. However, query optimization techniques adopted in the existing DBMSs cannot cope with the new challenges. Implementation of query languages has led to the process of query optimization. Little is yet published about this subject for OODB’s. The current proposals are limited to optimization using algebraic operators, and cite the development of fully operational optimizers as a future goal [5]. Furthermore, optimization techniques developed are mainly for traditional disk-based databases, where only a small part of the database could fit the fast main memory (MM). They are partly obsolete because of the progress in hardware and by new working environments, especially the powerful interoperable workstations with individual (personal) databases, and MM’s comparable to those of mainframes a few years ago. Also, MM has become reliable enough to support data for extended periods of time without any crashes. Hence, the interaction with the disk for reliability purposes can be much more infrequent, i.e., individual updates need not be committed to the disk. All this progress requires a revision of traditional database designs, but opens the way to performance that no traditional implementation could achieve. In general, as we know, the most important operation for a query is the join operation. A query typically involves a sequence of joins. To determine a good join order for the query, which is an important decision specified in a query execution plan (QEP).

In the following sections we introduce MOOD system, define cost functions, explain query processing in OODB and then discuss the retrieval operation in MOOD system. Finally we present our conclusion.

2. **MOOD System**

Main Memory Object Oriented Database (MOOD) is a new OODB system that will give better performance by eliminating the factors responsible for performance degradation. Many real time or time bound applications require high transaction throughput and limited transaction completion time. Disk resident object oriented database systems, however, cannot deliver the kind of high performance required by such applications, mainly because of their dependence on slow (as compared to memory) data storage devices such as disks. MOOD
achieves high performance by placing the entire database in main memory and uses a new high performance data structure that leads to faster data manipulation [4].

The approaches we used for MOOD are motivated by the characteristics of a MMDB and exploit the technology of stable memory in a genuine manner that differs from the numerous other proposals. To implement an object-oriented database in main memory the data structure consists of a multi-tiered hierarchy of classes, each of which contains a pointer to its parent super class, its first child subclass, and the next sibling class at its own level. For each class node, the name of the class is stored as a string variable along with a linked list holding the names of all of the class attributes. The attributes may be either primitive data types, such as integer or string, or objects of another class. There is also a linked list containing the specific instances of the class. These instance lists contain the actual data as opposed to the class hierarchy and attribute lists which would be considered to represent the database schema.

3. PERFORMANCE CONCERNS

Performance is a major concern for query processing as well as other component of the MMDB architecture. However, performance in this case depends primarily on processing time, and not on the I/O operations (magnetic disk or similar storage devices). This contrasts with models of disk-based systems which count I/O operations to determine the performance of an algorithm. Not only are the metrics different in MMDB analysis, but the architecture components are different as well. Minimizing the computation time of the algorithm is the most important factor to take into consideration when evaluating cost functions for a main memory database. In this paper we are ignoring issues involving network performance, congestion, etc., and therefore will be defining the communication time between client and server as a constant. However, it is possible that this may represent the largest component of the query execution time but nonetheless this is not the component of the cost function that we wish to focus.

One of the hardest problems in query processing is query optimization that requires accurate estimate of the costs of alternative query plans. Cost-based query optimizers assign an estimated "cost" to each possible query plan, and choose the plan with the least cost. Costs are used to estimate the runtime cost of evaluating the query, in terms of the number of I/O operations required, the CPU time, and other factors. The set of query plans examined is formed by examining the possible access paths (e.g. index scan, sequential scan) and join algorithms (e.g. sort-merge join, hash join, nested loops). There could be a large number of query plans depending on the complexity of the SQL query [7] requiring large search space. Generally speaking, it is not possible to examine all existing query plans hence; one optimization that is made is, to only consider a subset of the plans that are possible in theory.

Optimizers evaluate query plans using a mathematical model of query execution costs that relies heavily on estimates of the cardinality, or number of tuples, flowing through each edge in a query plan. Cardinality estimation in turn depend on estimates of the selection factor (a measure for how many objects will be retrieved for a given value of an attribute) of predicates in the query. Traditionally, database systems estimate selectivities through fairly detailed statistics on the distribution of values in each attribute. This technique works well for estimation of selectivities of individual predicates. However many queries have conjunctions of predicates such as, SELECT Count(*) FROM R, S WHERE R.Make='Honda' and R.Model='Accord'. Query predicates are often highly correlated (for example, model='Accord' implies make='Honda'), and it is very hard to estimate the selectivity of the
conjunct in general. Poor cardinality estimates and uncaught correlation are one of the main reasons why query optimizers pick poor query plans [7]. Do object-oriented databases or main memory databases create special issues when determining selectivities? This is one of the questions that will have to be addressed when deciding on an algorithm for MOOD system and will be discussed later.

Query processing algorithms and optimization techniques in disk resident database systems are based on the assumption that the data is disk resident. Therefore, in such cases the leaves of the query tree correspond to reading of the data from the disk, thus making number of leaves of a query tree a factor in decision making which leads to forcing query trees to have lesser number of leaves. For main memory databases however, this assumption is not valid and so we do not need to focus on this factor as well. Another factor related to query processing in a disk resident DBMS is the high cost of access to secondary storage, compared to the other factors (such as the algorithm used to join two or more objects, or the data structures in which indexes are stored, etc). Because of this high cost some simple query optimizers ignore other factors all together and compare different query strategies entirely on the number of data blocks that must be transferred between the disk and main memory. Thus standard query optimizing algorithms used for disk resident databases cannot be used for main memory databases. A cost function based entirely on the number of disk accesses would be completely meaningless for a system like MOOD. A cost function for optimizing queries in a main memory database must focus on the following functions;

1. **Storage cost:** This is the cost of storing any intermediate results that are generated by an execution strategy for the query. Most of the current approaches store these temporary results in main or secondary memory. This step is applicable to main memory databases as well since temporary results do still need to be created or stored. However, unlike disk resident databases, in a main memory database the space occupied by the temporary objects is in direct competition with the permanent objects for the space in main memory. Therefore, the goal should be to minimize the amount of space occupied by the temporary objects and if possible use some kind of highly efficient mechanism to store temporary object and to relate them to actual values.

2. **Processing cost:** This is the cost of performing in-memory operations on the data during query execution. Such operations include, for example, searching, sorting, merging and join operations, and performing computations on field values. In a main memory database systems like MOOD this will probably be one of the most important factors in the cost function.

3. **Memory usage cost:** This is the cost based on memory usage that pertains to the number of memory buffers needed during query execution. In a main memory database there really will be no distinction between this factor and the storage cost since the entire database is already main memory resident.

4. **Data Storage Structure**

This section describes the data structure that is being implemented in MOOD system in order to reduce the processing, memory usage and storage cost functions. The description of the data structure will help in understanding optimization and query processing algorithms that we have proposed for such a system.
When looking for reducing processing time, it should be obvious that one of the design goals for a query processing algorithm for a main memory databases would be to minimize the number of steps needed to find a specific data value in the database. In an object-oriented database the data is not organized as a set of relations but rather as a set of classes which are related to one another in a tree structure.

The features that are of interest here are the superclass, subclass and next sibling pointers. These pointers are part of each class in the class hierarchy. This architecture makes it possible to reach any class only by traversing a number of pointers equal to the depth of the class in the tree. Hence, while the number of classes in the database has the potential to expand exponentially the steps needed to find a given class would only increase proportional to the log of the number of classes, of course we are assuming a balanced tree. The data structure shown in figure 1, thus stores each object as a set of values stored in a contiguous chunk of memory. This could be implemented as a “struct” in any high level programming language, with each field of the structure holding a pointer to the actual data value.

Figure 2 shows the details of the record structure. In this record structure each field in the record is actually a pointer to a specific data value stored elsewhere. This is a good strategy for a number of reasons over storing the actual data value in the structure. First of all, since all fields are pointers, and for a given computer architecture all pointers would have the same length regardless of the data type to which they point, then all records for a given class would have exactly the same size in bytes. It would therefore be possible to locate a given attribute in an instance merely by knowing the offset of the appropriate field from the beginning of the record and following a single pointer stored at that location to the data value. Using this architecture the attribute list attached to the class would only need to store the name of the attribute and the offset, (either in fields or bytes) of the attribute from the beginning of each record. In contrast, if the actual data was stored in the record structure there would always need to be taken into the consideration that some data values (and instances) may require more space than others and this would considerably complicate the algorithms needed to locate these values.
Another advantage of using pointer structure is that each unique data value need only be stored once in the database. For a key attribute this will not make a difference as each object has a unique value for that attribute. But for non-key attributes the effect could be dramatic. Using this approach the amount of space that is used to hold the actual data can be significantly reduced, which while not necessarily speed up query processing, will nonetheless improve the performance of the database by allowing more objects to be stored in the given space.

Since the data structure shown is for object-oriented database the “values” shown in figure 2 may not be of a primitive data type such as the text strings shown in this example, instead may be complex objects. This fact makes it even more important that pointers, rather than the data itself, be stored in the object since objects can be quite large and should not be duplicated unless absolutely necessary. In fact, an object “data value” will be another record structure only in a different class. So this tends to even further reduce the amount of space needed to hold literal primitive data values. So we have addressed two cost functions namely, Storage and Processing in designing this data structure.

5. **QUERY PROCESSING**

Query processing for memory resident data must focus on processing costs, whereas most conventional query processing approaches attempt to minimize disk access [3]. However, processing costs can be difficult to measure in a complex data management system. Therefore, costly operations (e.g., creating an index or copying data) need to be identified first, and then strategies must be designed to reduce their occurrence. Operation costs may vary substantially from system to system, so that an optimization technique that works well in one system may perform poorly in another. Since sequential access is not significantly faster than random access in a memory resident database, query processing techniques that take advantage of faster sequential access lose that advantage [2, 3]. An example is sortmerge join processing, which first creates sequential access by sorting the joined objects. Although the
sorted objects could be represented easily in a main memory database using pointer lists, there is really no need for this since much of the motivation for sorting is already lost.

At the highest level a query is a statement written in a data manipulation language like SQL (DML) that describes in general terms what information should be returned to the user or what changes should be made to the data in the database. Before such a query can be optimized it must first be scanned by the scanner, parsed by the parser, and validated. The scanner identifies the language tokens such as, SQL keywords, attribute names, and object and class names, in the text of the query. The parser then checks the query syntax to determine if it is formulated according to the syntax rules of the query language. The query must also be validated, by checking that all attribute and relation names are valid identifiers as described in the schema of the database. The query is now available in an intermediate form and ready to be optimized for execution. Query optimizer is a component of the DBMS that attempts to determine the most efficient way to execute a query [7]. Figure 3 [6], shows all of the above mentioned steps of processing a high level query.

One of the most important steps in query processing is query optimization. The purpose of the optimizer is to translate this input query into a “query plan” (or query execution plan). In general, there are normally more than one ways to execute a given query, with widely varying performance. The query optimizer evaluates some of the different plans for executing the query, each of which would in itself be correct, and returns what it considers the most efficient alternative.

Most query optimizers represent query plans as a query tree of "plan nodes". This query tree is a tree data structure that corresponds to a relational algebra expressions where as, a plan node encapsulates a single operation that is required to execute the query. Each node has zero or more child nodes. These child nodes generate output that is then fed as input to the parent node. For example, a join node will have two child nodes, which represent the two join operands. The leaves of the tree are nodes which produce results by scanning the raw data (stored as objects) itself, through an index scan or a sequential scan [7].

The initial query tree that is created by the optimizer, and which is the starting point of the optimization algorithm, is said to be in “canonical” form. This consists of first joining all of the objects involved in the query using cartesian product, then applying select operations on the joined objects, and finally applying project operations on the results of the first two operations to output only the attributes of interest. An example of a query tree in canonical
form is shown in figure 4. This initial query tree can then be transformed into other equivalent trees using the general transformation rules for relational algebra operations. Each such transformed tree is then evaluated using an appropriate cost function to select the tree with minimal cost as the basis for the final execution plan.

6. **QUERY PROCESSING MOOD**

In general there are two types of queries; read only queries in which data values are retrieved but no changes are made to the database, and update queries in which data value is changed of the existing data, or deleted completely. These queries can be point queries or range queries. However, in case of OODB queries can be further classified as a single class or a class hierarchy. In the following sections we explain how the values for all the cost functions can be reduced by using a proposed query processing algorithm that is very effective and efficient for a system like MOOD.

6.1 **Read Queries**

In this paper we will only consider read queries since these involve search operations that are also part of the update query as well. We will defer the discussion about point, range, single class and class hierarchy queries to a later section. A read query (which is written as a “SELECT-FROM-WHERE” structure in SQL) can be considered to have three separate components, when broken down into its fundamental operations in relational algebra. This is commonly referred to as SPJ (Select-Project-Join). In a traditional disk-resident database, selection is often the first operation to be performed. There are number of reasons for this approach; to reduce the number of tuples as early as possible in query execution, it also reduces the size of the temporary results that must be created in main memory to hold the results created by the intermediate steps of the query. In addition it also reduces the number of disk blocks that need to be transferred into main memory, which is usually the most important issue for query optimization in disk-resident databases. Finally, since the physical storage of tuples and objects in disk files generally stores the values of all the attributes of a relation in adjacent memory locations, when retrieving a tuple or an object all of the attributes would be retrieved from the disk regardless of whether those attributes would later be used or eliminated by the projection operation, hence it is of no advantage to perform project before select operation.

\[\text{RESULTS} \]
\[\pi \text{(selection conditions)} \]
\[\sigma \text{(selection conditions)} \]
\[\text{k} \]
\[\alpha \]
\[\text{TABLE A} \quad \text{TABLE B} \quad \text{TABLE C} \]

*Figure #4: Canonical Form*
Where as revisiting the canonical query tree (figure 4) reveals that a query in canonical form is “Join-Select-Project”; first the objects involved in the query are all joined together as a cartesian product, then selection is performed to reduce the number of instances, and finally the desired attributes are chosen through projection. This is probably one of the worst query processing algorithms that you could possibly choose when it comes to efficiency but is conceptually easy to model. In processing a query in MOOD these operations are reordered to achieve a highly efficient query execution thus keeping the value for cost functions low. The operation that is performed first in this case is PROJECT. The detail about the entire query processing process is given in the following section after explaining the relationship between classes and queries.

6.2 Queries and Classes

In this section we would first explain data structure that is used in MOOD to store the database schema before explaining how queries will be executed. In a memory resident database it is not necessary to move any actual data values when executing a query. All that needs to be done is to create a new data structure for a class representing the results of the query. The attribute list for this new class will then be created by linking the attribute nodes from each of the already existing classes that are specified in the query. Since each attribute node contains the offset of that attribute in the records of its class, it is useful for project operation during query execution.

In the query processing algorithm for MOOD a new class is created that stores the results of the query. For a SELECT or JOIN operation a class for storing query results is created by retrieving the root node of the class (or classes) that contains the instances being sought and find the specific data records by using an index on the attribute specified in the query (if one exists) or scan the record list directly in the case where there is no index that can be used. The
attribute list of the class can have an index of its own based upon the name of the attribute that allows quick access to any attribute in the list. Selection of a specific attribute is required to evaluate conditions specified for select and join operations and also to perform project operations. Therefore the existence of an attribute name index is essential. mT-tree [10] being an efficient tree structure for implementing such an index will be used for the attribute list index for each class, as well as for all indexes for the data itself. mT-Tree combined with the MOOD data structure will reduce the number of steps needed to reach any specific instance (or set of instances) in the database. This helps in keeping the value for cost function low.

However, the results of the query, while modeled as a class, is not exactly the same as a regular class in the database, since it is not required to exist after the query results have been returned to the user. Therefore the data structure used to hold the data for a regular class does not need to be used for the query results as well. In fact, to facilitate the query processing algorithm a special data structure will be used to store query results in a new class as shown in figure 5. This class will not store actual data values so the rationale for performing selection before projection, used for disk resident database system, is not valid for MOOD. We propose that the first step for query processing in a MOOD should be the project operation on all of the classes specified in the query. This operation starts with finding each class specified in the query and for each class do the following steps:

- Create a new result class having structure shown in figure 5.
- Search the attribute list of the class, and for each attribute in the list that is displayed in the query results, append its attribute node to the attribute list of the new result class.
- If the attribute has the same name as another attribute that is to be projected (but differs only in its class name) set the “duplication” flag in the attribute node data structure.

In MOOD the class structure will include for each class an attribute list comprised of nodes. All of the actual data in the database will be stored in Data Nodes. Therefore all of the “P pointers” shown in figure 2 will be DataNode* in the case where the data is of a primitive type, or a Record* if the data is an instance of another class. Once each class has been found the next three steps are performed for that class. The main point for MOOD query processing algorithm is that: If possible, each class should be visited only once, thus reducing the processing cost. In addition this approach takes an advantage of the fact that project operations can be done with equal efficiency at any point in the algorithm and therefore perform the project operation and the select operations on a given class at the same time instead of waiting until the select and join operations had been completed before doing project operations as is usually the case with disk-resident databases.

Finally we look at select operations. One of the general transformation rules for relational algebra operations is that a conjunctive selection condition can be broken up into a cascade (that is, a sequence) of individual selection operations. Therefore any select conditions that involve only constant expressions of a particular attribute of the current class can be executed first time the class visited. This can be followed by the projection operation for the same class. Therefore the number of levels below the parent class at which the value of the select condition occurs will determine the order in which the select operations are performed. Selects with the smallest degree of level difference between the condition and the data records will be done first.
7. CONCLUSION

As database technology is applied to more and more application domains, user queries become more and more complex. The acceptance of OODB Systems required an efficient query processing algorithm. Since queries in an OODB could be very complex, query optimization for such queries becomes very challenging. Existing query optimization techniques either take too much time (e.g. dynamic programming) or yield a poor execution plan (e.g. simple heuristics). The quality of the execution plan generated for a given query is still unsatisfactory because these techniques do not take the special characteristics of a complex query into consideration. In this paper, we presented a further discussion of the query processing algorithm that is proposed for main memory object oriented (MOOD). In order to estimate the performance of this algorithm we presented cost functions that affect the performance of such a query processing algorithm. Finally we used these cost functions to conclude that the proposed query processing algorithm would be efficient. At present, we are studying the implementation issues of the proposed algorithm.

References:


Rapid Prototyping Data Integration

Barbara Lauridsen, National University Adjunct, Barbara.lauridsen@natuniv.edu

Abstract

Rapid Prototyping provides a tangible demonstration of system requirements with proactive user involvement. The overall goal of effective prototyping is to have no surprises during deployment of a large complex solution. New software development tools allow a prototype to be quickly modified to align with evolving perception of business requirements. The benefits of prototyping high-tech solutions can reduce development life cycle time and costs. This paper summarizes the advantages, pitfalls and cost effectiveness for rapid prototyping of user interfaces. It expands upon that theme to improve the interactive database that is to be redesigned. It addresses the opportunity to ensure quality and accuracy of the database content for an expanded software application. Illustrations and testimony from a real life data-driven prototype experience culminates the session’s learning opportunity.

Keywords: Prototyping, usability, Rapid Application Development, RAD, 80/20 rule, Storyboarding, Proof of Concept, Data Analysis Approach, Front-end, Back-end.

Rapid Prototyping gives non-technical users a tangible demonstration of system requirements. Software development tools allow a prototype to be modified to align with evolving perception of requirements. With the right team assembled, this method enables interactive recognition of business rules supporting metric-based evaluation of a potential system solution. “Software prototyping is driven by a confluence of different requirements not necessarily complementary to each other if not downright contradictory [1]”. The motivators are time-to-market, profitability for new products or upgrades to existing systems, streamlining the life cycle, leveraging legacy applications, maximizing web tools, reducing risk and sustaining market share.

According to Manor, rapid prototyping is “a software development process that allows usable systems to be built in as little as 60-90 days, often with some compromises [4]”. The 80/20 rule is often cited to mean that the most of the business requirements for a system can be fully satisfied even when some of its operational requirements are not. For the last ten years, high tech teams have been attempting to speed up the process by aiming at a result that meets expectations and will be funded.

Industry and trade publications on the topic frequently focus on web enabling traditional applications in order to activate web technology and open up access to a diverse customer base. However, a complete and accepted prototype of the user interface design is only a beginning. A complimentary design of the database is essential before the life cycle can be defined as complete. When complete, an outcome of a working prototype forms dynamic specifications of a solution accepted by user representatives. It is the milestone
accomplishment of an iterative design process. Whereas, a prototype might be a throw-away, stakeholders desire that the end result is not thrown away because it failed to implement a solution that works.

For this discussion, the term *front-end* defines user interfaces where information is revealed, whereas the term *back-end* defines the database structure where data is stored. A survey of the literature reveals that few publish articles about back-end prototyping, so, this conference paper offers ideas of a data resource integrator with in-depth experience in the field as a data modeler and data base administrator.

To demonstrate a real-life project which ensures the quality and performance of data to be modeled into an integrated solution, this conference session demonstrates a prototype of the back-end. The slide-show walk-thru reveals a prototype that integrated disparate local data into a central integrated regional database and automated data collection processes. The scope of this example project extended an existing application about habitats within a geographical region of forest service land. New data collection events involved sightings of endangered species in the area. Data about sightings became consolidated into an overall integrated database for the Natural Resources Information System.

The purpose of effective prototyping includes:

- A collaborative creative process for sharing ideas, define required functionality and user interfaces.
- An evaluation of the solution that can prove usefulness to the business.
- Increases confidence that all business requirements are known and doable.
- Communicates, promotes or supports design ideas.
- Defines or establishes selection criteria for technology with which to deploy the ultimate solution.
- Demo to give the customer an idea about what a product will look like.
- Recognition of the data model requirements.

Related concepts are: proof of concept, throwaway prototype, paper prototyping, video prototyping, storyboarding, demonstration prototyping and rapid tooling.

The potential benefits of rapid prototyping includes: a solution that reflects requirements that are known and understood, decreases life cycle time and costs, reduces potential for expensive design mistakes or oversights, constrains scope, impresses the customer with quick response, expedites delivery of software system. Backend prototyping increases quality of database design and loading strategy. The effort for gathering accurate data will not be underestimated.

**Advantages**

Rapid prototyping is a collaborative process focused on an outcome that allows a functional design team to identify navigation and usability problems to support a metric-
based evaluation of business activity. When a new or enhanced system is essential, it establishes a shared awareness with users of the business rules to be enabled by software and data that involves stakeholders who will be funding the effort to build a real solution. Normally, the outcome provides details about the user interface to the technical design team so they can build web pages and reports. Ideally, this limits a project's exposure to the forces of change [3].

Database prototyping aims at setting a quality standard, incorporating technical requirements quickly, in real time, and achieving early consensus about the scope of the back-end effort and its scheduled milestone to consolidate disparate source data into an integrated solution.

**Disadvantages**

Design features may be limited by the capability of the prototyping tool, the biases of the participants and exclusion of good design ideas. To prevent runaway schedules a Rapid Application Design (RAD) project needs a team already disciplined in time management.

**Cost-Effectiveness**

Although it is difficult to measure tangible return on investment, iterative prototyping usually demonstrates a time savings for securing the requirements and scope for building a complex system solution. Industry experience has shown that the method highlights any problems with system design so that design flaws or oversights do not impact the final software solution [7].

Formalized user testing of a software prototype proves usability of the front-end [2]. If realistic back-end data is included early, the outcome of a demonstration prototype of the look and feel interfaces can lead to confidence in allocating the resources to build the real software and then model the database to be loaded into the relational database.

**Considerations**

This usability argument endorses a method of brainstorming, creating, testing and communicating ideas about the system to be developed. The inspiration is to have a low-cost alternative to prove that a solution is viable during the early design phase to guide the high-cost building for complex software. Prototyping of the front-end accepts the importance of low-fidelity, low-tech media for usability testing with a focus on the user interfaces that lead to web page and report designs.

Special consideration should be given to accessibility of the system when the prototype is developed e.g. flexibility in choosing text size, font and color, good layout, operation with a screen reader. Accessibility should be planned for at the early prototyping stage even if implementation comes later [7].
Before participating in evaluation sessions, users outside of the design team are normally required to sign a formal consent to protect intellectual property of the concept and content. Data recorded must be kept confidential unless it is disguised.

Tradeoffs determine the pace of development. Efficient development balances economy, schedule, and quality aiming at faster than average, costing less than average and a better than average quality. The criterion for acceptance of deliverables must be fitness for a business purpose.

For RAD, something other than schedule must be negotiable. These will be the economy or the quality considerations. Resources allocation, scope, functionality included, bells and whistles and the 20% of the 80-20 rule can all be negotiated. For high priority projects there might be the formalized awareness of other projects to be displaced because people assigned to the team will be diverted. Time-boxing is a technique that limits wasting time talking about the exceptions. It is important to isolate secondary features that may be dropped in order to stay on schedule for the 80% normal functionality.

The process includes: brainstorming, paper or whiteboard exercises, software prototype with realistic and representative data, design specification for user interface, modeling of the data.

Materials & Tools for prototyping are a critical success factor. For low-tech prototyping: whiteboard, paper, pencil, adhesive notes, cards, scissors, etc. in a project room to test interactions. Snyder offers a comprehensive resource for understanding a simple but powerful low tech technique [6]. For high-tech prototyping: Collaboration tools, video conferencing, design and display tools for mockups, front-end software e.g. PowerPoint, Visio, WYSIWYG web development tool, visual basic or C, etc, plus back-end database engineering and loading, interfaces, etc [6].

Participants and stakeholders are essential to success. Active support of teamwork dynamics accepts the idea that high performance is voluntary and that teamwork provides a structure to contribute to a Rapid Prototyping project. The basic paradigm shift needed is to minimize the cost of failure measured by results and team momentum. The incentive to be involved is an important factor. According to Wilson (2003), "A highly functioning team does not need either carrots or sticks [9].” Teamwork offers alternatives such as synergy that occurs when those assigned become motivated to become a highly functioning team with a tangible result.

The cast of characters for a successful RAD project are: a trained facilitator, a hands-on project leader, business operations experts, change agents, customer representatives, design team members, graphics designers, specialists in using the tools, a data modeler, systems Administrator.

To avoid failure, a list of possible pitfalls of rapid prototyping projects is a handy reference [2]. The list includes:
• Focusing on the wrong things or with too much detail.
• Slipping into the 20% exceptions of the 80-20 rule.
• A basis for a quick hack product.
• Users becoming intimidated by an over developed prototype, withholding critique.
• Biased preferences about usability alternatives.
• Abandoning the next step, letting it become a sub-optimal final system.
• Paper prototyping won’t simulate identify some interactions features e.g. long documents, scrolling, rolling and cascading menus, response times.
• Software prototyping without realistic representative or with made-up data.
• Quickly conceived and poorly designed database structure haunting both programmers and users.
• Ignoring accessibility and performance issues.
• Relying on a desktop engine not designed to run within the internet.
• Since most demo prototypes are undocumented, only those who can maintain it are the people who originally built it.
• Forgetting that it is *only a prototype*.

This paper has summarized the opportunity to illustrate user interfaces but expanded the focus to achieving a *Data Driven* approach. The overall goal of effective prototyping is to have no surprises during deployment of a large complex solution both for front-end and for back-end.

For user interface (UI) testing, users may be observed and their dialog activity recorded to capture the way in which navigation between pages is done. Interviewing users or requesting electronic feedback is a source of insight about weak and strong features of the design. Severity of any problems can be identified that may justify calibrating the prototype as specifications for programming. Best is when the tool is flexible enough to perform revisions rapidly.

After end user testing of UI design, however, a project facilitator should allocate additional tasks for prototyping the interactions with the system’s backend to ensure a comprehensive database design and rules for data collection.

After UI acceptance, the next step is to extract the design implications, make recommendations for improvements and also clarify the scope for the technical development for UI programming, data collection and database design.

Iterative refinements and re-assessment of the requirements should occur until it is determined to be meeting usability goals and that the data modeling is focused on providing accurate and complete back-end. After confidence of the stakeholder has fully funded the effort, the phase commonly called “code-like-crazy” will quickly accomplish the promised savings because any risk of expensive re-work has been diminished.
References

   URL: http://www.sapdesignguild.org/community/book_people/print_review_eff_prototyping.asp
   Chapter 3 can be downloaded from portal: http://searchsoftwarequality.techtarget.com/tip/0,289483,sid92_gci1245581,00.html

   http://www.stcsig.org/usability/newsletter/0007-prototypingvisio.html


   Note: This web site presents a preview of the Usability Body of Knowledge (BoK) to guide practitioners in usability and user-centered design focused on the creation of representations of a user interface to a system.


Relational Database Loader for Implementing Advanced Queries in Protein Informatics

Sharen A. Eskelsen, Coleman College, Seskelsen@Coleman.edu
Scott S. Kelman, Coleman College, SKelman@Coleman.edu
Carl S. Ewig, Coleman College, CEwig@Coleman.edu

Abstract

The rapidly increasing numbers and complexities of the databases used in current genomics and proteomics research is often not fully exploited, due to limitations in the types of queries that may be posed. The form of the database should maximize the granularity, that is the number of possible tables and fields, in order to maximize the information that can be extracted. In this paper we present some initial development of a new form of parser/loader that takes all the proteomics data available in widely available text files and converts them into relational form that may be queried using all the features of SQL. We use as an example the SwissProt database. SwissProt is the most comprehensive and widely used database of protein properties as mapped to their defining sequences. We show results of querying the new relational database with queries that are often not possible with other available databases.

Keywords: Bioinformatics, Proteomics, Relational databases

Introduction

The databases used in life science research are rapidly increasing in both number and size. The number of databases currently publicly available and in common use has been estimated to be over 1,000 [1]. At the same time the sizes of most of these in terms of numbers of entries and also annotations are increasing nearly exponentially. This has led to the truism that for example a universal cure for cancer is probably already contained in existing databases somewhere. Researchers simply have not yet asked the right questions, in terms of complexity and specificity.

A number of approaches have been developed for attempting to extract more information from existing databases by using more complex queries. One is database integration and federation, in which combined items of data and information from several sources, which may be of multiple types, employing multiple formats or schema, and even in different physical locations, are retrieved as needed [2]. Another approach is to combine all the biological information into a single data warehouse for subsequent data mining [3].

Our approach is to create a software system (parser/loader) for building relational databases with arbitrary and adjustable levels of granularity, so that in principle it is not limited in the logical connections it can draw between any blocks of text in any combination of records. Search terms may be any field, character string, or even single letters. If the information is loaded into a
relational database, all the features of SQL may be used to define the logic of the queries. This may be contrasted with the use of nearly all publicly available search tools, which only allow searching for the presence of specified single keywords, sometimes connected by Boolean operators.

Construction of loaders for other types of applications, such as medical informatics, have been described previously [4]. Our approach differs in being directed toward implementing a highly flexible parser/loader with variable levels of granularity for defining searchable types of data, while exploiting the rich annotations available in modern proteomics databases.

Principles of the “Hypersearchable” Parser/Loader and Relational Database

For this project we selected the SwissProt [5] database. SwissProt is by far the most comprehensive and widely used database of protein properties as mapped to their defining sequences in terms of number of entries and number of annotated features per entry. The current release supports over 30 types of features mapped to specific positions in the amino acid sequences. These include, for example, splice variants, mutated sites, carbohydrate binding sites, lipid binding sites, metal binding sites, DNA binding regions, signal sequences, transmembrane regions, secondary structures, and active sites. Many of these are further qualified, for example allowing six different types of attached lipids, and nine possible post-translational modifications for a given sequence position.

The SwissProt database is made accessible as a single formatted text file. Our approach was to develop application software that can read in the data file while imposing an advanced level of granularity in terms of numbers, sizes, and types of the tables to be created, and use that information to create a relational Oracle database. An outline of the application is shown in Fig.1.

Construction of the Parser/Loader

The loader was written in the C# language on the .Net2005 platform. A table was created in an Oracle database for each selected type of data. The data type was located based on where it occurs in the text file. The SwissProt record accession number was employed as the primary key for each table. The file was read in by lines, and each field in the line was parsed into the appropriate column of its table.

Sample Queries

As examples of the use of this approach, we have constructed a relational database than can be searched for the following simple queries.

I. Search details of the entry text annotation of the protein sequences, as opposed to search fields such as organism name, by keywords.

   Example: Find all proteins that contain a charge relay system in their active site.

   Results: Charge relay triads are a complex but very common molecular feature of many enzymes, such as acetylcholinesterases and serine proteases [6]. Identifying this feature is
central to both determining protein function and designing inhibitors such as pharmaceuticals.
To find all proteins with this feature we search for “Charge relay system” in the ACT_SITE
feature. The results returned 5114 such structures, with as many as 3 per protein, each with its
identifying accession number for retrieving the complete SwissProt record.

II. Analyze database using queries that involve simultaneously both selecting and counting.
  Example: Find all proteins that have five or more disulfide regions in the same protein.
  Results: Disulfide or cysteine bridges are sulfur-sulfur linkages that are structural features
  of most proteins, imparting structural rigidity. For example there are three in insulin. A query
  using the DISULFID feature returned, somewhat surprisingly, 3031 proteins with 5 or more
  disulfide bonds. 362 proteins had more than 20. Each included its identifying accession number
  for retrieving the complete SwissProt record.
  However many of the protein entries with high numbers of disulfides are actually
  condensed or polymeric structures.

III. Search for logical combinations of details of the entry text annotations of the protein
  sequences.
  Example: Find all proteins that bind DNA and also contain disulfide bonds, but exclude
  all members of the kinase family of proteins.
  Results: The query found only 34 such structures, each with its identifying accession number
  for retrieving the complete SwissProt record.

IV. Any statistical analysis of the entries.
  Example: How many serine kinase proteins are known in homo sapiens and what is their
  average length in amino acids?
  Results: The number of known human kinase proteins in the database was 54 and their
  average length was 824.35 residues. Each included its identifying accession number for
  retrieving the complete SwissProt record.

V. Any alphanumeric data not present in specific fields or features.
  Example: Identify the protein isolated from Caenorhabditis elegans (round worm) with a
  molecular weight of 28163.
  Results: The molecular weight, along with the number of amino acids and the CRC64
  checksum appear in the header to the sequence data. Using this query alone, one unique protein
  from C. elegans is found with this the selected molecular weight. It corresponds to a protein-
  binding protein of the 14-3-3 family [7].

Conclusions

By employing properly constructed parser/loader software it is possible to generate relational
databases that can use nearly all the annotations and other textual data in the flatfile form of the
SwissProt database that combined with all the logical features of Oracle SQL. This
“hypersearchable” form can have arbitrary granularity, for example as large as whole protein
sequences or down to the level of individual alphanumeric characters. The ultimate goal is to
provide a bioinformatics research tool that will be publicly accessible over the Internet when
combined with an appropriate version of SQL for forming new types of proteomics queries.
Figures

Figure 1. Using the parser/loader application with a proteomics database to extend the ranges of SQL queries including fields, features, and qualifiers as well as the underlying amino acid sequences.

References

Categorization of Functional Dependencies for a Minimal Cover

Ajeet A. Chikkamanur
Department of Computer Science
REVA Institute of Technology & Management
Bangalore 560064 India
ac.ajeet@gmail.com

Shivanand M. Handigund
Department of Computer Science
Bangalore Institute of Technology
Bangalore 56004 India
smhandigund@gmail.com

Abstract

The database design of an information system is based on its business rules and functionalities abstracted from the requirements specification. The entities, their interrelationships, attributes of entities, and the functional dependencies amongst attributes of an entity, depend on the business rules buried in the application. The good database design principles require the preservation of these functional dependencies, but at the same time the increase in the number of functional dependencies enhances the complexity of the design. Thus, there is a need to reduce the complexity without sacrificing any of the functional dependencies. This is possible by the proposed automated technique that finds the minimal cover of the functional dependencies. Thus, the number of functional dependencies is reduced without compromising the design quality.

The paper proposes an efficient methodology for categorization of the functional dependencies. The algorithm achieves the goal, first by abstracting & retaining the absolutely necessary functional dependencies, then by discarding the unconditional implied dependencies & retaining the conditional dependencies for further processing. The conditional dependencies determine the presence of functional dependencies based on the order of their processing for the purpose. This saves the resource and time as it processes the necessity of each functional dependency in at most two accesses.

Keywords: functional dependency, hypergraph, Incidence matrix, categorization

1. Introduction

The database design for an application is based on its business rules and functionalities abstracted from the application. The entities, their interrelationships, attributes of entities, and the functional dependencies amongst attributes of an entity, depend on the business rules buried in the application. The good database design principles require the preservation and at the same time minimization of these functional dependencies. Thus, the number of functional dependencies and the design quality are inversely proportional to each other.

The database schema design commences with collection of data items and rules for dependencies amongst data items. The data items are represented as attribute of relation with restrictive conditions among them. These restrictive conditions are abstracted from the business rules and represented in the form of “functional dependencies”. The functional dependencies are assertion about the real world and hence the design of database is done without questioning them. From the set of functional dependencies, an efficient database is to be designed and hence we have to look out at any logical implication of dependencies such that the number of functional dependencies is minimal.

The paper proposes a methodology to categorize the functional dependencies in to three groups: absolutely essential functional dependencies, unconditional implied functional dependencies and conditional implied functional dependencies.

2. Background

This section presents the hypergraph nodes and edge representation from the functional dependency and Incidence matrix representation of hypergraph.

2.1 Functional Dependency

A set of functional dependencies represented by hypergraph [1] \( H = (N, E) \) over an attribute set \( A \) with number of vertices \( N = A \), and edge with following:

\[
E = \{(X, Y): \text{F}(X, Y) \in F \text{ and } Y \subset X\}
\]  

Where \( F (X, Y) \), with \( X \) and \( Y \) are subsets of \( A \), and uniquely defines the value of attributes in \( Y \). The set of attributes \( Y \) is not a subset or equal to the set of attributes \( X \).

A functional dependency graph [2] is a labeled graph with two kinds of nodes and edges, i.e. the dependency may contain one or two attributes on left hand side and one attribute on right hand side.
2.2 Incidence matrix

The hypergraph contains directed hyper edges. The directed hyper edge is represented as \( E = (T(E), H(E)) \), where \( T(E) \) is set of tail vertices and \( H(E) \) is set of head vertices. From the equation (1), the set of tail vertices \( T(E) = X \) and the set of head vertices \( H(E) = Y \). The incidence matrix of hypergraph \( H \) is \((V \times E)\) matrix and an element \( a_{ij} \) defined as follows

\[
a_{ij} = \begin{cases} 
-1 & \text{if } v_i \in T(E) \\
1 & \text{if } v_i \in H(E) \\
0 & \text{otherwise} 
\end{cases} \tag{2}
\]

3. Framework

A methodology for an abstraction of functional dependencies from a business rules is available [3, 9]. The application of good database design principles on these functional dependencies may enhance the quality of a designed database. This quality mainly depends on the existence of fewer dependencies. Thus, the main task in the database design is to obtain minimal cover of functional dependencies for the existing functional dependencies set.

Many methodologies proposed by various researchers [5, 6, 7, 8]. These methodologies suffer from some kind of deficiencies. For example, a database with \( n \) number of functional dependencies obtains the minimal cover by \( n^2 / 2 = O(n^2) \) accesses of functional dependencies. These numbers of accesses can be drastically reduced by determining absolutely essential functional dependencies, unconditional implied functional dependencies and conditional implied functional dependencies.

The absolutely necessary dependencies are to be retained in all the minimal covers and these dependencies are accessed only once. The unconditional implied functional dependencies are accessed only once and can be discarded. These need not be considered for other alternative minimal covers. The conditional implied functional dependencies are to be studied for their preservation in the minimal cover. This process is further refined, by altering the order of pair of conditional functional dependencies.

In an ideal condition, if almost one third of functional dependencies are conditional functional dependencies and complexity of accessing these functional dependencies is \( O(n) \). Moreover in practice the number of conditional functional dependencies is much less than \( n / 3 \). Therefore the order of accessing functional dependencies through the proposed methodology is always \( O(n) \).

For remaining less than one third of functional dependencies that are unconditional not absolute essential functional dependencies, the preservation is tested by altering the order of functional dependencies. The usual minimal cover methodology requires \( O(n^2) \) number of accesses but our proposed methodology requires \( O(n) + O(n^2/18) \). Thus for \( n \leq 18 \) the number of accesses reduces to \( O(n) \).

The methodology can be constructed by following steps:

1) Abstract the attributes and functional dependencies from the application domain.

2) Represent the attributes and functional dependencies by hypergraph in terms of incidence matrix representation. This step enhances the performance of methodology by eliminating trivial dependencies.

3) Categorize the functional dependencies into groups:

   a) Absolute essential functional dependencies
   b) Unconditional implied functional dependencies
   c) Conditional implied functional dependencies.

The first step of methodology abstracts the functional dependencies from the application and the methodologies for their abstraction are available [3, 9]. The functional dependencies are represented with canonical form i.e. the left hand side of functional dependency consist of one or more attributes and right hand side with one attribute.

The functional dependencies and attributes represented by the hypergraph in terms of incidence matrix representation. The rows and columns assigned with attributes and functional dependencies respectively. The edge of hypergraph consists of two parts: set of head vertices, set of tail vertices and functional dependencies consist of two parts: key attributes and non-key attributes.

For each key attributes in a functional dependency the entry for a hypergraph in terms of incidence matrix is \(-1\), for non-key attributes entry is \(1\) and \(0\) for non-existence of any attribute. The representation of hypergraph in terms of incidence matrix eliminates the trivial dependency and retains the key attribute of trivial dependency. This is a first step that will improve the performance of methodology i.e. all the trivial functional dependencies are eliminated and are not considered in the accesses.

The proposed algorithms categorize the functional dependencies in to three groups:

1) Absolute essential functional dependencies

   Some functional dependencies are not inferred from the other functional dependencies and essential in the set of functional dependencies i.e. the set of key attributes or its components of functional dependency cannot be inferred from the other functional dependencies.
2) Unconditional implied functional dependencies

Some functional dependencies are eliminated to retain the absolute functional dependencies i.e. the functional dependency implied from the absolute essential functional dependencies exclusively.

3) Conditional implied functional dependencies

Some functional dependency’s existence depends on the existence of another functional dependency. To retain one functional dependency, another functional dependency is dropped or vice-versa. It leads to two groups of conditional implied functional dependencies: One with set of functional dependencies by retaining conditional implied functional dependency and another set with ignoring the conditional implied functional dependency.

The algorithms for categorization of functional dependencies are shown below:

Input: Incidence matrix of functional dependencies
Output: Categorized set of functional dependencies

/* ALGORITHM FOR EXTRACTING ABSOLUTE ESSENTIAL FUNCTIONAL DEPENDENCIES */

fd = Initial functional dependency
while ( for all functional dependencies )
{
    for ( all attributes of functional dependency )
        /* single key attribute fd’s */
        if ( LHS attributes = 1 )
            fd’s with single key = fd;
    else
        {
            for ( i = 0 ; i < Number of fd’s ; i++ )
            {
                if ( LHS attributes = 1 )
                    next [ i ] = LHS attributes of ith fd
                if ( initial [ fd ] ∪ next [ i ] != initial [ fd ] & & fd != i )
                    { /* fd with multiple key attributes and their components are not existing in other fd */
                        Absolutely essential fd’s = fd;
                        break;
                    }
            }
        }

    /* fd which is to be tested for next category */
    To be test fd’s = fd;
}

/∗ ALGORITHM FOR EXTRACTING UNCONDITIONAL IMPLIED FUNCTIONAL DEPENDENCIES */
for ( all test functional dependencies )
{
    testl = tail attributes of test functional dependency;
    testr = head attributes of test functional dependency;
    repeat ( )
    {
        for ( all Absolutely essential functional dependencies )
        {
            lhs = tail attributes of Absolutely essential fd;
            rhs = head attributes of Absolutely essential fd;
            if ( testl ∪ lhs = = testl & & result fd not repeated ) /* broaden the functional base */
                testl = testl ∪ rhs;
            if ( testl ∩ testr = = testr ) /* test for trivial result */
                { delete test functional dependency;
                    return ( );
                }
            if ( Check for all Absolutely essential fd’s is completed )
                { To be tested fd = fd;
                    return ( );
                }
        }
    }
}

/* ALGORITHM FOR EXTRACTING CONDITIONAL IMPLIED FUNCTIONAL DEPENDENCIES */

implied ( )
{
    for ( i = 0 ; i < test functional dependencies ; i++ )
    {
        testl = tail attributes of test functional dependency;
        testr = head attributes of test functional dependency;
        repeat ( )
        {
            for ( j = 0 ; j < result and test functional dependencies ; j++ )
            {
                lhs = tail attributes of result fd;
                rhs = head attributes of result fd;
                if ( testl ∪ lhs = = testl & & result fd not used )
                    testl = testl ∪ rhs;
                if ( testl ∩ testr = = testr )
                    { swap = true;
                        temp1 = i° test functional dependency;
                        temp2 = j° test functional dependency;
                        delete i° test functional dependency;
                    }
            }
        }
    }
}
return;
}
if ( result and test functional dependencies fulfilled)
{ result = fd;
    return();
}
repeat();
}
if ( swap == true && swapping is one time with temp1 and temp 2 )
{ swap ( temp 1 , temp 2 ) in test functional dependencies;
    implied ();
}

4. Case study

To illustrate the methodology consider the following functional dependencies [4]

1) AB \rightarrow C  2) C \rightarrow A  3) BC \rightarrow D  4) ACD \rightarrow B
5) D \rightarrow E  6) D \rightarrow G  7) BE \rightarrow C  8) CG \rightarrow B
9) CG \rightarrow D  10) CE \rightarrow A  11) CE \rightarrow G

In the hypergraph in terms of incidence matrix representation the rows of a matrix are attributes and columns for functional dependencies. The key attributes are entered with -1, non-key attributes with 1 and 0 for non-existence of attribute for a functional dependency.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Incidence Matrix

The incidence matrix representation is depicted in figure 1. The trivial dependencies are eliminated when the functional dependencies are represented in hypergraph in terms of incidence matrix. For example consider a trivial dependency AB \rightarrow A, in which the key attribute A on left hand side of functional dependency, is entered with -1 value in incidence matrix and for right hand side attribute A is to be entered with 1 in the same position. Hence it is difficult to represent and is ignored.

The functional dependencies categorized in to three groups.

1) Absolute essential functional dependencies

The functional dependency categorized to this group by verifying the existence of key attributes and their components in another functional dependency. In functional dependency AB \rightarrow C the key attributes are AB and no other functional dependency consist of AB, A or B as their key attributes. So AB \rightarrow C categorized to absolute essential functional dependency. Similarly BE \rightarrow C belongs to this category.

The functional dependency D \rightarrow E has only one key attribute and there is no possibility of existence of its components. At the most there is a duplicate functional dependency or transitive dependencies. Many researchers [5, 6, 7, 8] provided the methodology for minimal cover of functional dependencies when key attribute is single and is applied to the set of functional dependencies with single key attribute. The methodologies are utilized for creation of this category. Hence, the functional dependencies D \rightarrow G, C \rightarrow A, D \rightarrow E are added to the category of absolutely essential functional dependencies.

2) Unconditional implied functional dependencies

The functional dependency, which is implied from the absolute essential functional dependencies, is dropped to retain absolute essential functional dependency. To determine implication of functional dependency, the base of key attribute is broadened by the essential functional dependencies and if it is resulted in a trivial dependency then the functional dependency is dropped.

In the functional dependency CE \rightarrow A, the base of key attributes CE broadened by substituting C by A (based on functional dependency C \rightarrow A). The result after broadening the functional base is ACE \rightarrow A. Since attribute A appears on right and left hand side i.e. trivial dependency and hence the functional dependency ACE \rightarrow A is dropped.

3) Conditional implied functional dependencies

The third category of functional dependencies is resulted from the functional dependencies neither belongs to absolutely essential functional dependencies nor unconditional implied functional dependencies. In this category the existence of one functional dependency depends on the existence of another functional dependency i.e. existence of both functional dependencies are leads to trivial condition and single functional dependency to nontrivial condition. To retain one functional dependency, another functional dependency is
dropped or vice-versa. This implication results in to two groups.

The functional dependency CG → B key attributes base is broadened by the C → A, CG → D and ACD → B i.e. ABCDG → B. The functional dependency CG → B resulted in trivial functional dependency by ACD → B. It leads to categorization into two groups. One with retaining CG → B and eliminating ACD → B and other with eliminating CG → B and retaining ACD → B.

When CG → B is retained the ACD → B, CG → D (C→A, CG → B and BC → D makes trivial functional dependency) eliminated and on other way when CG → B is eliminated then ACD → B, CG → D are retained.

The three sets of functional dependencies:

1) Absolute essential functional dependencies
   [AB → C, D→EG, C → A and BE → C]

2) Unconditional implied functional dependencies
   [CE → A]

3) Conditional implied functional dependencies
   a) By deleting CG → B and retaining ACD→B
      [BC → D, CG → D, CE→ G, ACD→B]
   b) By retaining CG → B and deleting ACD → B
      [BC → D, CG → B, CE→G]

5. Related work.

There are methodologies available [5, 6, 7, and 8] to obtain the minimal cover for the functional dependencies, where left hand side contains only one attribute. Few authors attempted to obtain minimal cover when left hand side contains more than one attribute [10, 11] and these methodologies are differ in their efficiency metric based on use of computational technique and hardware.

6. Conclusion.

An attempt is made to develop a methodology to categorize the functional dependency for a minimal cover, when left hand side has two or more attributes by reducing the functional dependencies size to nearly one third of the original dependencies size. Further, the proposal can be used for determination of minimal cover of functional dependencies and hence, to facilitate the normalisation process to normalise 2NF and 4NF. This may help in the development of automated normalisation proces.

References

CHOMS: A Portal for Obesity Management

Radha Nandkumar1, Arun Datta2, Jacqueline Caesar2, Weimo Zhu3, Peter Bajcsy1, Jeff Sunderlin4, Victoria Jackson4, Jill Sproat4, Heidi Krahling3

1National Center for Supercomputing Applications, 2National University Community Research Institute, National University, San Diego, 3University of Illinois at Urbana-Champaign, 4Illinois Department of Public Health

Abstract

Obesity, which leads to various weight-related health problems among children and adolescents, is continuing to be a growing concern in the United States. Development of effective surveillance, prevention, treatment and management strategies to address the health, social and emotional problems associated with overweight, particularly among school-aged children is very critical. Extensive research is already in place to address this issue and find suitable mechanisms to arrest this problem. The American Academy of Pediatrics (AAP) recommends that the Body Mass Index (BMI) should be measured on all youth as part of normal health supervision. BMI, which is easy to measure and correlates with body fat, assesses the weight status of an individual to identify those at risk and provide information for appropriate diet and physical activity.

The State of Illinois (IL) has initiated and legislated BMI surveillance and screening programs for all students attending public school systems. Legacy data associated with student health records and services are limited to manual paper-based processes. Under a multi-institutional joint collaborative pilot project, a web-based tracking system, termed Childhood Healthcare and Obesity Management System (CHOMS), is under development. This is designed for tracking and monitoring BMI as well as for efficient data management of childhood obesity in IL.

This online system is developed using open source software as a pilot project that will enable the users to collect data through multiple computerized means, store data in a relational database, and report related statistical and tracking information on youths at risk. CHOMS will provide web-based data entry forms for various users including healthcare service providers and help significantly in reducing manual paper-based work. Automatic computation of BMI percentile and the risk of obesity alert is embedded into this system. This system also determines and allows the user to view his/her health related report and takes into account the HIPAA compliance. For example, a student will be able to view his/her physical exam data, a nurse will be able to view healthcare service provided data, school official will be able to view school health report including immunization record of the student, etc. CHOMS will monitor the state of childhood obesity epidemic in IL and benefit school administrators to design and implement school-based obesity prevention and intervention programs. The system will also assist healthcare service providers in developing better services and guide communities as they undertake obesity prevention among children.
Why is C4I Software Hard to Develop?
Dr. Lee Whitt
Northrop Grumman Corporation
9326 Spectrum Center Blvd
San Diego, CA 92123
858-514-9400
lee.whitt@ngc.com

Abstract
Service Oriented Architecture (SOA) promises to transform the design, development, and deployment of C4I software, heralding a revolution in advanced and flexible warfighting capabilities….but it’s not going to happen, at least not for the next 10-15 years. Legacy C4I software, such as the Global Command and Control System (GCCS), will continue to prosper and evolve during this period, with the most visible “SOA revolution” consisting of point-to-point web services bolted onto legacy functionality….but don’t confuse this progress with the promise of SOA.

SOA technologies have been available for about 5 years now – a time frame that exceeds the threshold of patience for the next version of GCCS – yet viable SOA-based C4I systems remain elusive. So what’s the problem? The problem is that the truly hard issues of C4I are not being addressed, most prominently the complex business logic specific to C4I. By way of analogy, why is it hard to write software for stock market investors to select winners and avoid losers?

Developing C4I software is significantly more difficult than stock market software, because the business logic is far more complex. This complexity and the attendant implications are the focus of this white paper.

§1 Introduction

Many years ago, Ada was promoted as the silver-bullet programming language that would tame the C4I software beast – it didn’t happen. Then Java came along with its portability, mobility, and flexibility to wrestle C4I software into submission – it didn’t happen. Then the browser was offered as the fast path to success, delivering complex C4I capability with the simplicity of point-and-click access – it is still a work in progress, with only a few successes after 12+ years of intense development. Other past contenders for the silver-bullet technology award include C++, Common Object Request Broker Architecture (CORBA), Computer Assisted Software Engineering (CASE) tools, and Integrated Development Environment (IDE).

Most recently, Service Oriented Architecture (SOA) constructs, based on web services, have been anointed as the next technology silver-bullet for C4I. SOA governance is being institutionalized as the critical organizational/social component for success,
complemented by the artifact-generation processes resident in Capability Maturity Model® Integration (CMMI), Six Sigma, and DoD Architecture Framework (DODAF). All of these efforts have certainly created value, but they have not addressed the core challenge with the design and development of C4I software, namely the role of “business logic” in C4I. In particular, the C4I requirements process (defined and managed through organizational governance) has chronically failed to provide sufficient specificity for developers to deliver software that meets user expectations.

Today, we are besieged with SOA technology visions and marketing bluster, strongly influencing (even dominating) DoD thought and re-directing DoD resources into the transformation of legacy C4I software to SOA-based net-centric C4I capabilities. Indeed, the popular SOA headlines pervading Information Technology (IT) journals and web sites demand urgent action or face dire consequences:

- SOA – The Holy Grail of IT
- SOA isn’t optional – it’s imperative
- SOA - Start using it today or risk losing everything tomorrow
- SOA – Ignore it at your own peril
- SOA - The silver bullet to reduce costs, improve agility, and fast-track the delivery of products and services
- SOA – The pixie dust that makes everything marvelous
- SOA is sweeping through organizations, upending the competitive order
- SOA will completely re-define IT, providing infinite flexibility at the speed of light
- SOA is really quite simple and very powerful….and it will quickly fix all that ails your broken IT systems
- SOA is a career choice – don’t be unemployed

Fear seems to be a key motivator in the SOA discourse and the effect is corrosive in creating an environment for a balanced approach to defining the core issues in the design and development of C4I systems. Even a discussion on what’s easy and what’s hard in building C4I systems would be useful; there’s no balance in the SOA mantra that proclaims everything is easy, if only a few common mistakes are avoided. To help the reader calibrate his thinking relative to this paper, it is suggested the reader take a moment to mentally compile a list of what is relatively easy and what is truly hard in the design & development of C4I systems.

As we clamber to create the next-generation of C4I capabilities, it is not prudent to confer SOA technology on every C4I mission area and hope that a C4I mission solution will emerge. Technology is not a solution – only a solution is a solution – and a solution requires that data be processed, managed, and analyzed, under the tight control of well-defined rules (equivalently, business logic). SOA is not about defining complex C4I business logic and, so far, organizational SOA governance has not addressed the necessary details.
Today, legacy C4I software has been stigmatized and development efforts marginalized, while new SOA-based software initiatives are urgently (and sometimes recklessly) launched with the aim of quickly replacing legacy systems. On the surface, the strategy appears reasonable and straightforward, buttressed by a cavalcade of technology standards from industry groups and product offerings from SOA vendors. Some SOA evangelists (representing prominent SOA vendors) may even beckon with pronouncements that web services can be developed in only a few days and entire SOA systems completed in just a few months. Case studies provide ballast by documenting real-world SOA successes, where new capabilities were rapidly delivered at reduced cost and risk. For C4I, the blaring message is that most (all?) legacy software can be expeditiously replaced, often in less than 18 months, once the defense industry gets serious and sponsors commit the necessary resources.

So what is holding us back? SOA technologies have been around for over 5 years and the DoD commitment to SOA has been manifest and material. Given this and given the putative ease with which SOA-based systems can be created and delivered, it is reasonable to ask:

Where (in the battlespace) have SOA-based C4I systems replaced legacy C4I systems?

The stubborn fact is that there has been no replacement….and replacement is the litmus test - not running legacy systems in parallel, not maintaining legacy systems on the back-end, and not keeping legacy systems for “insurance”. The success of SOA prototypes in demonstrations and exercises is no substitute for real-world systems providing real-world capabilities.

So in spite of the great SOA shopping spree, the harsh reality is that legacy C4I software continues to thrive in the operational environment and, as argued below, will continue to be successful for another decade or more. Certainly, technology advances continue to provide a powerful engine for progress and innovation, but the core engine behind today’s C4I capabilities is the implementation of rule-sets: Complex, often highly-coupled, technology-agnostic, and context-dependent on the mission. As such, it will be very difficult to capture and replicate these rule-sets in a new system, and particularly difficult in an SOA-based system explicitly designed to mitigate the dependency on mission context, system state, and inter-process coupling.

This paper will focus on the business logic embedded in legacy C4I software and make the case that this logic is the key to successful C4I. It is the expression, implementation, and synchronization of this logic, developed over many years (sometimes 10 years or more), that is responsible for the effectiveness of today’s C4I systems. Ignoring the critical role of this logic will reduce SOA-based systems to thin shells for basic data transfer and distributed processing, with legacy C4I systems continuing to provide the “heavy lifting” through the processing, management, and analysis of tactical data.

§2 Background
The business logic in C4I applications takes many forms, so it will be helpful in this discussion to provide some examples. Perhaps the most obvious example is correlation logic, specifically the rule-set responsible for analyzing and correlating track data received in various forms (e.g., GOLD, TDIMF, TIBS, M-Series, J-Series, COP messages)\(^1\) from various sources/sensors (e.g., Radar, Sonar, EW, JSTARS, GPS, TADILs, overhead)\(^2\). Correlation logic is responsible for comparing the new track information contained in an incoming track report against a database of tracks and, based on various scoring factors (e.g., attribute “weights”, source confidence, correlation thresholds, user configurations), make a correlation decision. The decision can take various forms, such as:

- Correlate to existing track
- Create a new track
- Create an ambiguity, along with a list of correlation candidates
- Ignore/discard the report

Furthermore, the correlation logic may spawn secondary correlation effects, such as the automatic merging of two existing tracks into a single track. The simplicity of this discussion belies the remarkable complexity inherent in track correlation logic.

Another example of business logic is the deconfliction of forces, such as the deconfliction of subsurface activities (e.g., submarine movements and surface ship towed array operations) and the deconfliction of airspace activities (e.g., manned & unmanned flight missions, weapon fire zones, controlled ingress/egress routes). Creating and managing these activities in space and in time, while achieving mission goals and accommodating dynamic re-planning in response to unexpected events and conditions leads to a complex rule-set for deconfliction and operational safety.

Yet another example is the business logic embedded in security data guards, designed to inspect data “objects” at various levels (e.g., individual data fields, collection of data fields, metadata) and then take an action to sanitize the data object. The action can be as simple as making no change to the data or as invasive as excising portions of the data and replacing other portions with suitable data substitutes. The diversity of security enclaves and the diversity of data exchange agreements across enclave boundaries, particularly in coalition, multi-lateral, and bi-lateral operations, leads to companion complexity in the rule-sets….and the software implementation of the rule-set must be validated through a comprehensive and rigorous certification process.

In many legacy systems (such as the family of Common Operating Environment (COE) based C4I systems), the embedded complex rule-sets are not amenable to easy

\(^1\) Acronym definitions: Tactical Data Intercomputer Message Format (TDIMF), Tactical Information Broadcast Service (TIBS), M-series are LINK 11, J-series are LINK 16, Common Operational Picture (COP).

\(^2\) Electronic Warfare (EW), Joint Surveillance and Target Attack Radar System (JSTARS), Global Positioning System (GPS), Tactical Digital Information Link (TADIL)
decomposition and distribution across the network in the form of web services, for reasons that are inherent to the operational context (not bound to the design of COE-based systems), as will be discussed in this paper. Furthermore, many of these rule-sets are inter-related and operate in concert to support mission planning, execution, and situational awareness across the battlespace. This dependence engenders an additional level of complexity that likewise defies decomposition and distribution in the form of services.

§3 A Relevant Perspective on the Role of Business Logic

Consider the challenge of writing software designed to optimize an investor’s stock portfolio. The high-level system requirements are simply stated:

1. Identify stocks with a high probability to increase in value
2. Compute the maximum expected value and time frame to achieve it
3. Generate “sell” alerts whenever a stock falls below a computed threshold (computed based on a high probability the stock will decrease in value)

For this investment software project, assume that all economic information, market conditions, company performance information, etc., are readily available in real-time and in a well-defined format (e.g., XML)…..and assume further that there are no technology obstacles in terms of software tools, standards, architectures, etc. Under these conditions, will it be hard to create the investment software?

By our assumption, it will be straightforward to collect streams of information on markets and economies, and to write software to perform routine manipulations (e.g., extract and compare elements), but it is not at all clear what rule-set is needed to “intelligently” process and analyze the information to accurately predict the behavior of individual stocks. It’s not even clear how to a priori test the rule-set for consistency and completeness, nor how to verify the software faithfully implements the rule-set. A highly modular design will make it easier to build and modify the rule-set, but the path to success depends on deep domain expertise and disciplined software engineering.

In a similar fashion, it is relatively straightforward to collect streams of information across the battlespace, and SOA technologies will likely improve collection, support routine processing, and facilitate distribution. But C4I systems are responsible for advanced information processing and analysis, supporting complex mission planning and predictive battlespace awareness. C4I domain expertise, combined with detailed mission requirements and disciplined software engineering, is the key to defining and implementing viable mission rule-sets.

From a systems engineering perspective, C4I planning/execution and investment planning/execution have many similarities in the collection and processing of information, and subsequent analysis for decision-making. Hence, at a design level, SOA approaches for C4I software can be re-cast into SOA approaches for investment software,
with the exception that C4I software is (arguably) more difficult to design and develop – not because of technology, but because of the underlying business logic. The motive of an investor is well-defined and linear, characterized mostly by greed with elements of risk mitigation. The motives of today’s enemies are complex and highly non-linear (perhaps even chaotic), defying simple characterization. In this respect, the business logic for C4I software will necessarily be more complex than investment software.

Figure 1 is a recent Office of Naval Research (ONR) slide on a candidate information integration framework for C4I planning and decision-making. Replace the reference to “Warfighter” with “Investor” and the references to “Platforms, Sensors, Weapons, Sources, People” with “Industries, Financial Data, Marketing & New Products, Analysts, Consumers & Economic Activity” (respectively) and the result is an information integration framework for investment planning and decision-making.

![Figure 1: ONR slide showing a candidate Information Integration Framework for C4I planning and decision-making](image)

The business logic of C4I, and specifically the implementation of this logic, is the core engine that is responsible for the effectiveness of C4I systems. Get the logic wrong (represented by the yellow boxes in levels 2 & 3 of Figure 1) and the system will be degraded, perhaps rendering it unusable. It is the central importance and pivotal role of business logic – surpassing any other system characteristic, technology, or interface – that define and differentiate software applications. Of course, business logic must be consistent with policy directives and organizational goals so that, for example, investment logic can be tailored for different types of investment strategies (value investing, growth investing, income investing, etc.).

In summary, business logic is the genome of effective software. This perspective, along with the prodigious complexity in crafting, implementing, and deploying C4I business logic, is the central theme of this paper and the case for why C4I systems are hard to
build.....and by extension, the case for the longevity of legacy C4I systems. The following sections provide amplifying details.

§4 Edge-Cases, Uncertainty, and Ambiguity

Whenever business logic is designed, edge-cases, uncertainty, and ambiguity must be accommodated and managed. Data elements often carry inherent uncertainties and data sets may contain internal conflicts that lead to unexpected and ambiguous situations. Managing these situations is non-trivial and can have a profound effect on the viability and reliability of the business logic.

In C4I and financial investment, edge cases and ambiguous conditions often present the predominant challenges when trying to craft a complete and consistent rule-set....and highlight the conundrum that one person’s edge-case event is another person’s mainstream event. For example, how do we judge a negative report by one investment analyst on a particular stock – a failure to see the bigger picture or an insight into a trend? In C4I, does an enemy maneuver represent a deception or a pattern of activity? In general, there is no well-defined methodology for defining business logic to handle the diversity of conditions that occur at the boundary. Domain experts can disagree on the nature of a condition (boundary or mainstream), what it means, and how to respond, as evident to anyone who has watched financial experts disagree over stock picks or military experts disagree over tactics.

So what is a reasonable approach to designing C4I (or investment) business logic that can accommodate a broad range of known and unknown conditions, replete with uncertainty and ambiguity? Perhaps the optimal approach is to mitigate decision errors, specifically to balance two types of statistical errors:

- Type 1 Error (commission error – doing something that is incorrect)
- Type 2 Error (omission error – failing to do something that is correct)

The balance is an uncomfortable one. Type 1 errors lead to bad system decisions that occur at computer processing speeds, yielding a high volume of problems. Type 2 errors require user intervention to complete a process or task that the system failed to finish, yielding an overload of manual tasks queued for user action. In both cases, these errors can quickly degrade system utility and usability, erode user confidence, and lead to delays in decision-making and/or poor decisions.

§5 Context, Scalability, and Certification

All software processing occurs within a context. Sometimes the context is shallow requiring no knowledge or management of the current system state, but more meaningful processing requires a context. The example of track correlation is instructive, since an incoming track report cannot be processed in the absence of a track database, as the current state of this database is the context within which the correlation logic operates (cf., §2). Similarly, deconfliction logic requires a comprehensive battlespace context of
unit activities (equivalently, a database of unit positions, movements, mission assignments, and mission dynamics) in order to deconflict new mission activities.

A track correlation service, devoid of an underlying track database upon which to operate, would be ineffective at best. In an SOA environment, a user (or system) request to the correlation service to process and correlate a new track report would need access to the user’s (or system’s) current track database in order to effectively compute. Context is everything or, to paraphrase Vince Lombardi’s credo, “Context is the only thing”. Legacy systems do many things well, but they excel at managing system “state” and processing data “in context”….because tightly-coupled systems are well-suited for state management and context monitoring. SOA designs cannot achieve success in this area because the tenets of loosely-coupled design and dynamic discovery are anathema to preserving system state and operational context.

It is easy to expose data on the network through web services, but the underlying business logic (and associated context) responsible for the creation and fusion of data is not exposed and, hence, cannot be accessed by client applications. As a consequence, these clients have freedom of action in terms of actions and manipulation of the data, including data display, secondary fusion, data routing & distribution, etc. As a concrete example, if an F/A-18 flight route is created by mission planning software that faithfully models F/A-18 Hornet flight characteristics (e.g., turn rate, climb rate, maximum flight time), then it is a simple matter to create the flight route as an XML document defined by the Common Route Definition (CRD) schema and post it to a pub/sub repository for discovery and retrieval by applications. However, once an application pulls an XML-formatted flight route, there is no guarantee that the application will respect – or even know - the supporting business logic in the original mission planning application responsible for creating the route. If the route is modified by a (de-coupled and stateless) client application, then the resultant may be an “illegal” F/A-18 route (illegal according to the original rule-set) which could be re-submitted back into the pub/sub repository for other applications to discover and retrieve. In extreme cases, the route change might have the F/A-18 stopping in mid-air, flying into an obstacle, or remaining airborne beyond its fuel capacity.

The decoupling of data from applications that “own” them – without pedigree or reference back to the application and without constraint on allowable operations on the data – will lead to a significant degradation in data reliability and loss of confidence by users. Returning to one of the main motivations for object oriented programming (perhaps even the key driver for object oriented designs), the object construct bundled data with functionality, so that applications pulling data “objects” would also implicitly pull the underlying functionality responsible for managing the data. A worthy goal to be sure, but it had limited success in practice for many reasons (beyond the scope of this paper), but suffice it to say that much of the object’s embedded functionality was designed for accessing specific data elements and for data presentation, not the responsible business logic. Today, the replacement of data “objects” by XML documents as the lingua franca for data sharing translates into the replacement of data functionality
Scalability issues are closely related to context, because the more extensive the context, the more relevant the scalability to guarantee acceptable system performance. In the case of COE-based systems, recent advances in track correlation have expanded the track database to “unlimited”, meaning that there are no constraints on the size of the track database or its composition (i.e., the number of tracks of different “types”, such as LINK, ELINT, Acoustic, and Missile). Given the requirement to manage global track databases, exceeding several hundred thousand tracks and with update rates on the order of several hundred per second, it is imperative that correlation be operationally efficient and highly reliable. In this setting, it is unlikely that a context-free, distributed, loosely-coupled web service will offer much value in satisfying the performance and state management requirements for correlation over a broad range of data feeds, from (soft) realtime combat systems to non-realtime sources.

C4I legacy software has been generally developed to support a pre-defined client/server environment, meaning that the number of clients connected to servers is typically constrained (e.g., perhaps several hundred clients in a large installation). Indeed, efficiencies have been employed to optimize performance in this client/server environment, such as the use of the UDP and multi-cast protocols. In contrast, web services must be able to support thousands of clients, and efficiencies will be more difficult to achieve, particularly for web services that are bolted onto legacy software, as scalability will be constrained by both the web service and the backend legacy system. Since legacy business logic remains firmly entrenched in legacy systems, these scalability issues cannot be resolved without significant architectural changes.

Software certification is the test and evaluation of software to verify that it will perform as designed within the target operational environment (i.e., context) and subject to the expected data inputs and outputs (i.e., scalability). Legacy systems (such as the GCCS family of systems) generally have a well-defined certification process because the operational environment and input/output channels are well-defined and constrained. SOA designs, which permit dynamic discovery of services, ad hoc composeability of services into functional workflows, and unbounded data sources and data rates (e.g., via pub/sub repositories) are no longer subject to traditional constraints……and indeed may be unbounded. The difference between a good idea and a bad idea is that a good idea has bounds.

In this respect, as noted previously, it is not feasible to simply bolt web services onto legacy systems and expect the resultant SOA systems to be well-behaved. Furthermore, certification of legacy systems does not seamlessly yield certification of the associated “attached” web services nor does it confer certification onto ad hoc capabilities comprised of dynamically discovered web services.

The next-generation SOA enterprise will likely be a thin shell through which legacy systems continue to perform the “heavy lifting”, based on their embedded, complex – and
§6 Interoperability

Interoperability is the holy grail of C4I systems. Unfortunately, interoperability is rarely defined with precision (assuming a precise definition is even possible), though we seem to share an instinctive understanding of what it means – we’ll know when we don’t have it.

Much progress has been made over the last 10-15 years to field interoperable C4I systems. Perhaps the premier exemplar of interoperability is the COE-based family of systems, which has achieved an unparalleled record of success in delivering compatible C4I capabilities to US and coalition forces since 1995. These systems, and the underlying business logic, have been refined through years of operational use and direct interaction with the warfighter.

However, the emergence of SOA constructs for easily deployed and discoverable services breaks this traditional system engineering approach while creating new challenges (and perhaps set-backs) in furthering C4I interoperability. The problem is not with SOA technology standards and the attendant technical interoperability (defined by published specifications). Rather, it is the ease with which new business logic can be hosted in a net-centric environment and exposed as a discoverable web service….discoverable in terms of the interface requirements, but without discovery (or guaranteed interoperability) of the underlying business logic.

The tight controls imposed on the development of legacy C4I systems (e.g., COE-based systems), and the long test cycles - many will reasonably argue too long - prior to deployment, have been largely responsible for ensuring system effectiveness and deep interoperability across the battlespace. The paradigm shift to web services as rapidly-deployed independent components (independent of a larger C4I context), subject to compliance with SOA standards and approval to operate (e.g., connect to the network), exposes many new entry points that, if unchecked, will erode C4I interoperability. Over time, the proliferation of web services will result in many similar services, similar in terms of advertised functionality (discovered via UDDI and WSDL), but governed by different embedded rule-sets…..and “SOA Governance” is not a magic elixir, in spite of the rhetoric.

An example may help clarify this matter. In military exercises, some of the exercise data may be synthetic, intended to simulate real-world units and mission activities, while other data is real-world (e.g., real-world units reporting their real-world positions). In COE-based systems, there is a clear distinction between synthetic data and real-world data (identified by a “flag” in each track structure), preventing synthetic reports from updating real-world positions and vice versa. Furthermore, synthetic tracks cannot be merged/combined with real-world tracks to form a single track. Without arguing the rationale for this business rule, all COE-based systems provide a consistent
(interoperable) foundation for managing and distributing exercise and real-world data across the global network.

In the future, it is conceivable that circumstances will arise when a community of interest (COI) will decide that it is acceptable to blur the distinction and separation between exercise and real-world data, perhaps allowing selected tracks to be merged. Certainly, a web service that performs the merger could be easily developed and quickly deployed, based on an agreement (within a COI) as to its intended use. But once the web service is registered on the network, it can be discovered by a larger community of users (and processes), perhaps leading to usage beyond the original agreement - the unintended consequence.

This example represents a simple case, but the fundamental concern is the emergence of inconsistent and incompatible business logic, built into discoverable web services, that control the processing and analysis of tactical information for decision-makers. We already see this situation with the deployment of web services, provided by financial institutions, to help customers monitor and manage their investment. Different rule-sets, operating on the same market information, often lead to different investment recommendations, even for the same investment strategy. Furthermore, efforts to compute an “average” recommendation by aggregating and smoothing the recommendations from different financial institutions may not yield a viable investment strategy (e.g., it is not reasonable to “average” a buy and sell recommendation for the same stock).

To be clear, there is value in developing different sets of business logic, based on policy, goals, or other factors. In the last century, mathematicians actively studied game theory (cooperative and non-cooperative games) with a focus on developing optimal strategies. It was hoped that these strategies could be successfully applied to economic models (as rule-sets that capture the role of dominant economic factors), in the same way that calculus was successfully applied to physical models (e.g., computing planetary orbits based on the law of gravitation)....and of course, different models will yield different results.

In the context of C4I, it is a provocative question to ask whether interoperability trumps the rule-set (or model). Restated, the question is whether it is better to have a shared understanding of the battlespace based on an incorrect rule-set or to have different (and divergent) understandings of the battlespace based on better, but different, rule-sets. The core premise of the COP – and the reason for it being called “common” – is that interoperability trumps the rule-set. The reasoning is that it is better to improve a single rule-set than to deconflict and converge different/divergent rule-sets (and iterate this process for every new rule-set that emerges).

For SOA, the success of interoperability at a technical level (e.g., standards compliance) will be off-set by the failure of interoperability at a system level, as conflicting business logic is implemented in web services and hosted on the network for discovery and use. Governance may slow the proliferation of conflicting web services, but it will be
generally ineffective at forcing convergence to a common rule-set or limiting the use of conflicting web services. The resulting SOA environment will jeopardize the decade-long trend toward improved cross-service and coalition C4I interoperability, which began in 1995 with the release of the DII COE.

C4I interoperability is a deep concept, penetrating far below the technologies that permeate the veneer of system interfaces, data formats, and network protocols. C4I interoperability at a global level requires a shared, consistent methodology for data processing, management, and analysis across all systems.....and this means shared business logic across the network.

§7 Coalition Operations and Security Boundaries

One of the most difficult problems in C4I is seamlessly bridging security enclaves. Senior US military commanders have repeatedly asserted that all future military engagements will involve significant coalition forces, so we must be able to seamlessly share information, mission plans, and C2 orders across coalition boundaries. The US Navy’s Numbered Fleet Commands (C2F, C3F, C5F, C7F)\(^3\) annually collect and prioritize their top 10 operational C4 requirements to drive the acquisition process. For FY07, coalition and multinational C4I interoperability remained the top priority.

Today, the C4I data exchange environment is relatively simple, consisting mostly of strongly formatted messages, such as GOLD and TADIL messages, and various generic data sets, such as email, images, and database records, often with associated metadata tags. Data guards operate on the data payloads to sanitize the contents, according to an embedded rule-set that complies with a security policy specific to each boundary between two enclaves. Although data guards are improving, the development and test process remains very long, with little allowance to rapidly respond to a dynamic coalition battlespace.

Given the current state of data guards, the introduction of web services and the supporting infrastructure (e.g., UDDI, WSDL, BPEL, WS*)\(^4\) presents an overwhelming raft of new challenges. The ad hoc composeability of web services in realtime to create mission capabilities and workflow, coupled with the flexible transport protocols, SOAP handshake protocols, and formatting requirements of XML documents cannot be \textit{a priori} defined. Even the seemingly simple case of sanitizing the contents of an XML document require that the data guard have access to the XML schema and the dictionary for every field to ensure that a legal XML document is produced.

\(^3\) Acronym definitions: Commander Second Fleet (C2F), Commander Third Fleet (C3F), Commander Fifth Fleet (C5F), Commander Seventh Fleet (C7F)

\(^4\) Acronym definitions: Universal Description, Discovery, and Integration (UDDI), Web Service Definition Language (WSDL), Business Process Execution Language (BPEL), Web Service Security, Policy, Reliable Messaging, Addressing, etc. (WS*), Simple Object Access Protocol (SOAP), eXtensible Markup Language (XML)
The diversity of enclaves and the lack of controlling doctrine/policy to completely and consistently define the rules for processing XML documents, accessing (or replicating) UDDI, WSDL, PBEL, etc. across security boundaries presents challenges that will defy our best efforts through the next decade. This is not a failure of organizations or people, but rather a consequence of the complexity inherent in defining rule-sets, while ensuring sufficient flexibility to support SOA deployments, complying with organizational doctrine/policy, and achieving certification through rigorous and comprehensive testing of the implementation.

A cogent argument can be made that it will not be possible to build and deploy SOA-tempered data guards, given the dynamic nature of SOA and the expected spate of new/evolving SOA standards, web services, and XML schema. Instead, the state-of-the-art will be represented by minimalist data guards in specialized environments, defined by support for a limited number of static web services and XML schema.

**§8 Why is it hard to develop C4I capabilities**

In this section, the issues discussion above will be threaded into a perspective on the challenges in developing C4I capabilities, along with recommendations on how to address these challenges.

There are no easy solutions to the hard problems in C4I and most of the easy problems have already been solved and implemented in software (meaning that we understand the business logic for doing most of the relatively simple things in C4I, such as message processing, attribute-based track correlation, simple data visualization on a map, basic data distribution on a network, and attribute-level semantic interoperability). Indeed, the evolution of C4I systems over the last 2 decades (the period of this author’s involvement) has been astonishingly lackluster, faring much worse than the evolution of computer technology. During this period, advances in software technology have not yielded comparative advances in the C4I applications that leverage this technology. By contrast, advances in hardware technology have been stellar, sometimes besting the initial promise and yielding computer products that exceed expectations.

The slow progress in C4I software is not a technology shortfall but, as argued above, a reflection of our inability to specify the business logic for effectively processing and analyzing all of the battlespace information needed to address complex C4I requirements. The same argument applies to the slow progress in financial investment software, and the following “challenge” items apply equally to both C4I and investment domains.

- **Challenge - Business Logic**: Today, we cannot articulate C4I requirements with sufficient granularity to create comprehensive rule-sets for information processing and analysis that accommodates the broad range of battlespace information. This failure is manifest in the generalities found in current C4I requirements, leading to inconsistent interpretations and incompatible implementation. If there are any “grand challenges” in C4I, a detailed specification of requirements would be on the list.
**Recommendation:** Governance oversight must be expanded in several areas of the requirements definition process:

1. Most obvious, requirements should be expanded to provide more details. Recognizing the impossibility of doing this completely and consistently, even modest progress would be significant. Of course, the requirements process should generally avoid the common pitfall of specifying implementation details, but the SOA juggernaut makes this difficult.

2. System engineers (the ones responsible for translating the requirements into rule-sets and managing the implementation) should be involved in the requirements definition process. Generally, C4I domain experts craft the requirements, but they are not system engineers...in the same way that an airline pilot is a domain expert in all aspects of flying commercial aircraft, but he is not an aeronautical engineer.

3. Governance should mandate the inclusion of (proposed) test plans to clarify requirements and to assist with system verification....as part of the requirement specification. If the statement of a requirement is too generic for a system engineer (or a system tester) to define a material test plan, then the implementation will suffer from unwanted/unexpected behavior. If the buyer can’t define the criteria for acceptance in precise and unambiguous terms, then caveat emptor.

4. Governance should recognize the importance, prevalence, and diversity of edge cases and anomalous conditions, and developers should not implement type 1 or type 2 errors in code without clear guidance, including test cases.

- **Challenge - Preserving Context and State:** SOA designs are context-free and stateless, allowing any web service to be dynamically swapped with an equivalent web service located anywhere on the network. As discussed previously, this design is not always optimal or desirable.

Furthermore, business logic is often context dependent, so specific rule-sets (or subsets) become active when specific conditions are met....and segregating business logic from context and state may be problematic. As an example of the coupling between rule-sets and context/state, consider the challenge of managing bandwidth, specifically optimizing the use of available bandwidth. Bandwidth management tools generally rely on IP addresses and port assignments, along with the quality of service flags in the packet header, for routing and flow decisions. Each packet is managed independently of other packets, relying only on the information in the packet header.
With the emergence of IPv6, better bandwidth management tools will be available, but none of these tools have a deep contextual understanding of the data payload, such as the relevant mission areas, the business logic responsible for generating the information, and the relationship and priority vis-à-vis other packets on the network. Is a data stream on threat missile tracks a higher priority than a data stream on threat submarine activity? If the bandwidth is shared with missile defense systems, then the answer is easy. If the bandwidth is dedicated to a ship engaged in anti-submarine warfare and if the threat submarine is in the vicinity of the ship, then the answer is easy. Without access to context, bandwidth management tools will make suboptimal decisions. Furthermore, if some of the data is already available to the end systems (e.g., through organic sensors), then the data streams can be down-sampled, providing only the portions needed by the end system.....but this requires information on the state of each end system. The optimization of bandwidth requires a complex combination of business logic, context management, and state management. Technology alone will not suffice.

**Recommendation:** C4I system design must allow flexibility in technology decisions; SOA is not the best approach to every C4I requirement, particularly those where context and state are indispensable. In fact, many C4I requirements have context and state dependencies, leading to the following recommendations:

1. Governance should mandate that requirements include a description of the applicable operational context or system state. If there is no inherent context or state, then a statement to that effect should be included with the requirement.

2. In the definition of test plans (as described in the previous recommendation), the plan should accommodate context and state, if necessary.

3. To the extent possible, business logic should be defined and segregated according to context and state (recognizing that much logic may be shared across context and state boundaries).

- **Challenge - Interoperability:** The perennial challenge is to ensure interoperability across C4I systems. Interoperability with technical standards is relatively easy to achieve, but interoperability of rule-sets has been extraordinarily difficult to achieve, except when the rule-sets are shared through the use of a common implementation (as in the case of the COE). From a governance perspective, the core issue is whether consistency trumps flexibility. If the former, then interoperability will be advanced; if the latter, then interoperability will be elusive.

**Recommendation:** For complex software modules, it is generally better to fix problems and re-factor the software to support new technology, rather than
starting from scratch (with due recognition that there are exceptions). This approach preserves the embedded business logic and avoids lengthy (and costly) test and fix cycles. Of course, the re-factoring process should address scalability for network environments and efforts should be made to mitigate any new certification requirements….easier said than done!

Even if we could solve many of these hard problems in the evolution of next-generation C4I systems, an even deeper problem looms. In science, we look for patterns of behavior that can then be abstracted into principles (e.g., laws of physics, mathematical equations). This process allows us to capture the lessons learned of scientific research in a well-defined form. In software development, lessons learned are often captured in design patterns, derived from routine and repeatable software development tasks. The abstraction of these patterns into general constructs allows them to be used – and reused – in many settings. Examples of design patterns include the Model-View-Controller (MVC) pattern for Graphical User Interfaces (GUIs) and the Create-Retrieve-Update-Delete (CRUD) pattern for database interaction. The same approach for capturing patterns applies to CMMI and Six Sigma, where governance and quality control patterns are abstracted and broadly applied to the management of software development. The DODAF portfolio of Operational Views (OV), System Views (SV), and Technical Views (TV) provides various design patterns for architectural views of a C4I system.

Unfortunately, we have not yet been able to identify general design patterns for C4I mission rule-sets….and this in spite of decades of work and volumes of lessons learned. Perhaps we haven’t framed the issues correctly, or looked hard enough, or perhaps this area does not lend itself to patterns and abstraction. If the latter case, then we can expect C4I business logic to continue being hand-crafted by teams of domain experts and system engineers working long hours over many years. In this case, C4I systems will continue to be hard – very hard – to develop and advances in technology will be little help.

§9 Concluding Remarks

C4I program managers, stakeholders, and sponsors spend too much time on technology, architecture, and user interfaces, while ceding the core software tasks (e.g., responsible for data processing, aggregation, and analysis) to coders who often have minimal operational experience, minimal domain expertise, and often minimal supervision when coding C4I business logic. Anyone can write code to implement functionality. Anyone can write interfaces to capture data. But very few have the domain expertise (operational and engineering) to create comprehensive rule-sets and even fewer to effectively implement complex rule-sets into a C4I system.

This is not a technology issue and seeking technology-focused solutions will squander resources and fail to deliver on the promise. The introduction of SOA will not change these dynamics, and improved governance will have little impact unless it drills deeply into the engineering tank and provides guidance at a level of detail that is meaningful to coders. Indeed, the development of C4I business logic is a staggeringly complex and
time-consuming endeavor under the best of circumstances, such as found within the well-defined, pre-defined, and constrained environment of today’s legacy systems.

In developing C4I capabilities within an SOA environment, we seem fixated on the mechanics of creating atomic functions (i.e., web services), managing them through governance, and gluing them together via an Enterprise Service Bus (ESB) and workflow constructs. The real value in C4I capabilities springs from the business logic that fuels the evolution of data-to-information and information-to-understanding in order to achieve mission context, to predict battlespace trends, and to assess decision consequences.

Lack of attention to the preeminent role of C4I business logic will be largely responsible for the continued dependency on legacy systems over the next 10-15 years……and the business logic embedded in these legacy systems will tenaciously resist efforts to extract, decompose, and reconstitute the rule-sets into an interoperable SOA environment (interoperable with legacy systems at the deep level of business logic). During this period, promises to transform C4I through an infusion of new SOA technology will be costly, problematic, and delayed, though strong governance will help mitigate the easy problems.

Fear must be attenuated as a key motivation for rushing into SOA, and high-frequency marketing bluster like “Are you agile or fragile? Are you leading or failing? Are you growing or fading?” do not advance the discourse. The rigors of responsibility must temper the stampede to deploy SOA. Yet, in spite of the lack of real-world success in deploying SOA-based C4I systems over the last 5 years, we are harried into immediate action to embrace – are you ready for this - the next wave of SOA technology (called SOA 2.0, or Web 2.0 or maybe Enterprise 2.0), affording even more benefits and efficiencies…..please make it stop!

So how do we avoid the excesses of SOA thinking and moderate the hype surrounding SOA technologies? It is a profound mistake to view SOA as a solution, especially while ignoring the hard problems in C4I discussed in this paper. For every contrarian view of SOA, there are hundreds of briefs, white papers, webcasts, conferences, case studies, etc. extolling the virtues & benefits of SOA. SOA success in areas such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), payroll, and supply chain are possible because organizations have a reasonably firm understanding of the necessary business logic, with sufficient granularity for implementation. This is not the case with C4I as can be seen by reading any Request for Proposal (RFP) for a C4I system. We can avoid some of the sensationalism of SOA by focusing more attention on what the software is suppose to do rather than the mechanics of how it will be done.

Of course, progress in SOA will continue to be made, but near-term expectations for advanced SOA-based C4I capabilities will remain unfulfilled. Instead, legacy C4I systems will continue to be the load-bearing foundation for C4I in real-world operations, with point-to-point web services bolted on to evoke the success of SOA.
Dr. Lee Whitt is currently a Northrop Grumman Technical Fellow involved in the design and development of C4I systems. He is one of the original developers of the first desktop-based C4ISR prototype called JOTS (Joint Operational Tactical System), initially deployed in 1985 on Carrier Battle Groups and Ashore Command Centers in all theaters of operation. Throughout the 90s, he focused on the Pacific Theater, supporting ashore and afloat commands with the evolution of C4I systems from JOTS to JMCIS to GCCS. He led the development of the first web-based C4I system (ELVIS = Enhanced Linked Virtual Information System) in 1995 and the first Java-based C4I system (ELVIS II) in 1996. More recently, his interest is on the challenges of developing network-centric capabilities.

He earned his Ph.D. in mathematics from Yale University (1975) and worked in academia for several years as an assistant professor on the faculty of Cornell University and Texas A&M University.
Waveform Distortion Estimation during Multiple Power Quality Events

Amritpal Singh  
Dept. of Electronics and Communication Engineering  
Lovely Professional University, Phagwara, Punjab, India  
Email: amrit_lpu@hotmail.com

Mannohsan Singh  
Dept. of Electrical & Instrumentation  
SLIET, Longowal, Sangrur, Punjab, India

Abstract: Identification and classification of voltage disturbances in power system is an important task in power system monitoring and protection. This paper presents the characterization of voltage disturbances waveforms obtained from the simulation of 5-bus power system using MATLAB/SIMULINK 6. Feature extraction has been done using various digital processing techniques such as RMS, Filter Banks and FFT. This paper presents a comparison between different signal processing techniques and establishes that the dyadic filter bank analysis is most suitable technique for automatic characterization of various power quality events.

Keywords – Power Quality events, Classification, Digital Signal Processing Techniques.

I. Introduction

The deregulation [1] of power industry and the development of equipments sensitive to power disturbances have resulted in the increased competition in the power market. Consumers now expect uninterruptible supply. Hence it has become utmost important to take appropriate corrective action at right time to mitigate the after effects of different unwanted events such as symmetrical & unsymmetrical faults, line switching, transformer energizing, etc., if happen in the network. The protection system should efficiently isolate the faulted portions of the power system in case of faults but not in case of other events like transformer energizing, large motor starting, line switching, etc. And for this, the events have to be identified. For the identification, it is a very general approach to extract some features, which are unique and quite distinguishable for the events, which are to be identified and classified in different classes.

Further the development of sophisticated electronic equipment sensitive to power system disturbances [2] poses a stringent requirement of clean power. Many power system events particularly voltage disturbances are responsible for mal-operation of equipment like computers, process controllers and adjustable speed drives. Voltage disturbances in a power system can be caused by any power system event viz., faults or switching of line or transformer energizing etc. These voltage disturbances have been unanimously recognized as the most troublesome events capable of affecting commercial, industrial and household electric loads. Even very short voltage disturbance can provoke irreversible damage to sensitive equipment and impose significant economic losses due to unexpected interruptions of industrial production processes. These voltage disturbances can be broadly classified as voltage sags, voltage swells and voltage interruptions [3]. Towards the direction of intelligent power quality monitoring and development of automatic classification and analysis tools, appropriate signal processing tools are required in order to extract the information from the signals. The work in this paper is carried out along the following two lines:

1. Simulation of the power system under consideration provides the necessary data. With focus on voltage events viz., voltage sags, voltage swells and interruptions.
2. The use of signal processing tools for the extraction of the distinctive features of power system events mentioned earlier. The signal processing methods adopted are RMS, FFT and Filter banks.

II. Data Collection

Data analysis presented is on voltage measurements. Data presented and analyzed come from the simulation of 5-bus power system [4] as shown in Fig. 2.1. The simulation is done using the MATLAB/SIMULINK Power System Blockset.

Table 2.1: Data of the 5-Bus Network

<table>
<thead>
<tr>
<th>Bus Code (p-q)</th>
<th>Impedance $Z_{pq}$</th>
<th>Line Charging $Y_{pq}/2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>0.02 + j0.06</td>
<td>0.0 + j0.030</td>
</tr>
<tr>
<td>1-3</td>
<td>0.08 + j0.24</td>
<td>0.0 + j0.025</td>
</tr>
<tr>
<td>2-3</td>
<td>0.06 + j0.18</td>
<td>0.0 + j0.020</td>
</tr>
<tr>
<td>2-4</td>
<td>0.06 + j0.18</td>
<td>0.0 + j0.020</td>
</tr>
<tr>
<td>2-5</td>
<td>0.04 + j0.12</td>
<td>0.0 + j0.015</td>
</tr>
<tr>
<td>3-4</td>
<td>0.01 + j0.03</td>
<td>0.0 + j0.010</td>
</tr>
<tr>
<td>4-5</td>
<td>0.08 + j0.24</td>
<td>0.0 + j0.025</td>
</tr>
</tbody>
</table>
III. Power system simulation using MATLAB/ SIMULINK PSB

The use of computer simulation tools is essential in power system studies. Software tools are widely used by utilities for transient event simulations, power flow studies, stability analysis and operational planning. Most of the commercial available packages are designed to work with large power system models. The use of such tools is often cumbersome and not well suited to the small power system studies. The latest Power System Blockset (PSB) of MATLAB, a powerful graphic tool that allows building schematics and simulation of power systems. The blockset uses MATLAB/Simulink environment to represent common components, machines and devices found in electrical power networks. One of the important features of PSB is its ability to simulate either with continuous variable time step integration algorithms or with a discretized system. One of the advantages of MATLAB over other available softwares is its processing power and its ability to display the results while simulation is running. MATLAB also allows the user to perform complex post-processing on simulation results. Diagrams can be assembled simply by using click and drag procedures into Simulink windows. The Power System Blockset uses the same drawing and interactive dialogue boxes to enter parameters as in standard Simulink blocks. Simulation results can be visualized with Simulink scopes connected to outputs of measurement blocks available in the PSB library. These measurement blocks acts as an interface between the electrical blocks and the Simulink blocks.

The generators are simulated with a Simplified Synchronous Machine block. The lines are simulated with distributed parameter lines block. The figure shows the fault simulated on bus 1. The faults and 3-phase circuit breakers are simulated with blocks from the three-phase library. Various oscilloscopes are used at different measurement points to display voltage waveforms. Simulation results obtained for a single line to ground fault on bus 1 are shown in Fig. 3.2.

Fault is simulated at \( t = 300 \text{msec} \) and is cleared at \( t = 800 \text{msec} \). Fig. 3.3 Voltage on all buses during L-G Fault on bus 1

Bus 1 experiences a severe voltage dip of phase A and there is rise of the voltages of phase B and phase C during the fault period of 500msec. At Bus 1, the phase A peak voltage magnitude before fault is 1.0245 p.u. It falls to approx. zero during the fault which recovers to 1.0217 p.u. after the fault. The phase B pre-fault, during fault and after fault peak voltage magnitudes are respectively 1.0245 p.u., 1.7042 p.u. and 1.0511 p.u. at Bus 1. The phase C pre-fault, during fault and after fault peak voltage magnitudes are respectively 1.0659 p.u., 1.6923 p.u. and 1.0489 p.u. at Bus 1. Other buses also show change in voltage magnitude before the fault, during the fault and after the fault.

Fig. 3.2: Voltage disturbance waveforms during SLG fault on bus 1

In the same way, voltage disturbance waveforms for different events viz., symmetrical and unsymmetrical faults, transformer energizing and line switching are obtained from the simulation of the 5-bus power system under consideration. Transformer connected to bus 4 is energized at \( t = 800 \text{msec} \). The output voltage waveforms at the transformer input terminals are recorded and is as shown in Fig. 3.3

Fig. 3.3: Voltage waveforms on bus 4 during transformer energizing

With the increasing amount of measurement data from power quality monitors, it is desirable that analysis, characterization, classification and compression can be performed automatically.

3.1 Root Mean Square Method

RMS can be evaluated over a cycle or a half cycle window and is given [9] by

\[
s_{\text{rms}}(n+N) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} s^2(n+i)}
\]

Where, \( N \) is the window length.

The advantage of this method is its simplicity, speed of calculation and less requirement of memory because rms can be stored periodically instead of sample per sample. But its dependency on the length of the window is its disadvantage. Also rms does not distinguish between fundamental frequency harmonics or noise components, therefore accuracy will be more with less harmonics and noise components. Moreover, it is suitable for analyzing any event in a power system other than harmonics and noise. RMS values, continuously calculated for a moving window of the input voltage samples, provide a convenient measure of the magnitude evolution, because they express the energy content of the signal. The basic idea is to follow.
the voltage magnitude changes as close as possible during the disturbing event. The more RMS values are calculated, the closer the disturbing event is represented, especially the non-rectangular variations. Fig. 3.4 shows the voltage at bus 1 and its corresponding RMS plot. The fault is Single line to ground fault at bus 1.

![Voltage and RMS plot at bus 1 during SLG Fault at Bus 1](image1)

There is a severe voltage dip of phase A caused by the Phase A to ground fault at bus 1 which is revealed by the RMS plot of the voltage at bus 1. Also there is voltage rise in both phase B and phase C. It is seen that after the fault is cleared the voltages of all the three phases takes some time to recover to the pre-fault voltage. Following parameters are evaluated in this method to characterize the voltage disturbances due to the above mentioned events:

### 3.1.1 Maximum and Minimum rms Magnitude

The root mean square (rms) value of the voltage waveform is calculated using the following equation

\[
    s_{\text{rms}}(n+N) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} s^2(n+i)} \quad (3.1)
\]

Where, N is the window length.

Based on equation (3.1), rms values of the voltage signal during the whole event duration are computed, from which the following two parameters are obtained.

(i) **Maximum rms value**

The highest rms value of the waveform during the sample duration.

(ii) **Minimum rms value**

The lowest rms value of the waveform during the sample duration.

### 3.1.2 Other obtained parameters

(i)**Voltage Disturbance Initial Time**

This refers to the moment when the rms value below 0.95 p.u.

(ii)**Voltage Disturbance End Time**

This refers to the time when the rms value recovers to within 0.96 p.u.

(iii)**Disturbance Duration**

The difference between the disturbance initial and end times.

(iv)**Initialization Angle**

The angle at which the sag occurs using the positive-going zero crossing as the reference angle.

(v)**Recovery Angle**

The angle at which the disturbance ends using the positive-going zero crossing as the reference point on wave.

(vi)**Initial Phase Angle Shift**

The phase angle shift is defined as the difference between the phase angle of the sag waveform that of the reference waveform. The reference waveform is defined as the 1.0 p.u. nominal waveform in synchronism with the pre-event waveform. It can be calculated by either the crossing method.

(vii)**Zero Crossing Method**

The zero crossings of the actual waveform compared to those of the reference waveform. phase shift is defined as the difference between zero crossings of the two waveforms.

(viii)**End Phase Angle Shift**

The calculation method for the end phase angle shift is similar to the initial phase angle shift except that the angle is calculated at the disturbance end moment.

(ix)**RMS Magnitude Unbalance Ratio**

This variable describes the unbalance degree between the three phase voltages during the sag. It is defined by equation[10]

\[
    \text{Unbalance ratio} = \frac{V_{\text{high}} - V_{\text{low}}}{V_A + V_B + V_C} \quad (3.1)
\]

where, \( V_{\text{high}} \) is the highest rms voltage of the three phases, \( V_{\text{low}} \) is the lowest rms voltage of the three phases, \( V_A \), \( V_B \), and \( V_C \) are rms voltages of the three phases respectively.

These parameters were calculated and are tabulated.

### Table 3.1: Parameters at Bus 1 (SLG fault at Bus 1)

<table>
<thead>
<tr>
<th>Disturbance Parameter</th>
<th>Phase A</th>
<th>Phase B</th>
<th>Phase C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum rms (p.u.)</td>
<td>1.1363e-005</td>
<td>0.6950</td>
<td>0.5928</td>
</tr>
<tr>
<td>Maximum rms (p.u.)</td>
<td>0.7682</td>
<td>1.2301</td>
<td>1.2859</td>
</tr>
<tr>
<td>Average rms (p.u.)</td>
<td>0.4506</td>
<td>0.9130</td>
<td>0.9351</td>
</tr>
<tr>
<td>Final rms (p.u.)</td>
<td>0.7257</td>
<td>0.7256</td>
<td>0.7256</td>
</tr>
<tr>
<td>Starting time (ms)</td>
<td>318.6</td>
<td>313.1</td>
<td>316.2</td>
</tr>
<tr>
<td>End time (ms)</td>
<td>802.3</td>
<td>810.9</td>
<td>806.5</td>
</tr>
<tr>
<td>Duration (ms)</td>
<td>483.7</td>
<td>497.8</td>
<td>490.3</td>
</tr>
<tr>
<td>Recovery time (ms)</td>
<td>21.3</td>
<td>48.5</td>
<td>47.2</td>
</tr>
<tr>
<td>Initialization angle (deg)</td>
<td>-74.85</td>
<td>-142.73</td>
<td>146.13</td>
</tr>
<tr>
<td>Recovery angle (deg)</td>
<td>-55.97</td>
<td>-139.7</td>
<td>136.13</td>
</tr>
<tr>
<td>Initial Phase angle shift (deg)</td>
<td>17.05</td>
<td>32.25</td>
<td>25.24</td>
</tr>
<tr>
<td>End Phase angle shift (deg)</td>
<td>56.64</td>
<td>16.71</td>
<td>132.36</td>
</tr>
<tr>
<td>Maximum rms unbalance ratio</td>
<td>2.9953</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Remarks

| Interruption | Swell | Swell |

Fig. 3.5 shows the voltage and its RMS plot on the output of the transformer which is connected to bus 4 is energized at \( t = 600\text{msec} \).
As seen from the Fig. 3.5 the RMS plot shows a very small voltage dip in all the three phases at bus 4 when the transformer is energized at t = 600msec.

Table 3.2: Parameters at Bus 4 (Transformer Energizing)

<table>
<thead>
<tr>
<th>Disturbance Parameter</th>
<th>Phase A (p.u.)</th>
<th>Phase B (p.u.)</th>
<th>Phase C (p.u.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum rms</td>
<td>0.5332</td>
<td>0.5236</td>
<td>0.5361</td>
</tr>
<tr>
<td>Maximum rms</td>
<td>0.5445</td>
<td>0.5429</td>
<td>0.5428</td>
</tr>
<tr>
<td>Average rms</td>
<td>0.5332</td>
<td>0.5236</td>
<td>0.5380</td>
</tr>
<tr>
<td>Final rms</td>
<td>0.5346</td>
<td>0.5271</td>
<td>0.5376</td>
</tr>
<tr>
<td>Maximum rms unbalance ratio</td>
<td>0.0394</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Remarks: Normal Normal Normal

3.2 Filter Banks

Low-pass, High-pass and Band-pass filters can be used to extract signals in specified bandwidth. The filter banks have been used to study in detail a specific band of frequencies.

Dyadic Analysis Filter Bank

The Dyadic Analysis Filter Bank block decomposes a broadband signal into a collection of successively more band-limited components by repeatedly dividing the frequency range. The typical (asymmetric) n-level filter bank structure is shown below in Fig. 3.6.

Scale 1: $f_1 = 1250$ Hz
Scale 2: $f_1 = 625$ Hz and $f_2 = 1250$ Hz
Scale 3: $f_1 = 312.5$ Hz and $f_2 = 625$ Hz
Scale 4: $f_1 = 156.25$ Hz and $f_2 = 312.5$ Hz
Scale 5: $f_1 = 78.125$ Hz and $f_2 = 156.25$ Hz
Fundamental: $f_1 = 39.0625$ Hz and $f_2 = 78.125$ Hz

Where, $f_1$, $f_2$ = lower and upper cutoff frequencies

The simulation is performed at the fault initiation as well as fault termination. Only initiation point calculations are made. The termination point calculations are out of the scope of this paper. The voltage plots in all these scales at bus 1 and 4 are shown in the following figures.

Table 3.3: Parameters at bus 1 during line-to-ground fault at bus 1

<table>
<thead>
<tr>
<th>Scale</th>
<th>Filter</th>
<th>$V_A$ (p.u.)</th>
<th>T msec</th>
<th>$V_B$ (p.u.)</th>
<th>T msec</th>
<th>$V_C$ (p.u.)</th>
<th>T msec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\approx 1250$Hz</td>
<td>0.001</td>
<td>300.6</td>
<td>0.001</td>
<td>301.2</td>
<td>0.001</td>
<td>300.6</td>
</tr>
<tr>
<td>2</td>
<td>625-1250 Hz</td>
<td>0.002</td>
<td>300.3</td>
<td>0.0028</td>
<td>300.5</td>
<td>0.001</td>
<td>301.4</td>
</tr>
<tr>
<td>3</td>
<td>312.5-625 Hz</td>
<td>0.005</td>
<td>301.8</td>
<td>0.002</td>
<td>300.4</td>
<td>0.00</td>
<td>301.9</td>
</tr>
<tr>
<td>4</td>
<td>156.25-312.5 Hz</td>
<td>0.018</td>
<td>300.4</td>
<td>0.012</td>
<td>301.3</td>
<td>0.013</td>
<td>300.5</td>
</tr>
<tr>
<td>5</td>
<td>78.125-156.25 Hz</td>
<td>0.012</td>
<td>300.3</td>
<td>0.028</td>
<td>300.4</td>
<td>0.023</td>
<td>301.6</td>
</tr>
</tbody>
</table>

3.3 Fast Fourier transform

A Fast Fourier Transform (FFT) is an efficient algorithm to compute the discrete Fourier transform (DFT) and its inverse.

Let $x_0, ..., x_{n-1}$ be complex numbers. The DFT is defined by the formula
A 128-point FFT is taken over half cycle running window with an overlap of 50 samples between each window during both fault initiation and fault termination.

**Window-1:** Half cycle window at fault initiation
**Window-2:** Half cycle window just after fault initiation with an overlap of 50 samples with window-1.

The method of Fast Fourier Transforms is applied to get the frequency response of the voltage disturbance waveforms for following events on buses 1 and 4 of the 5-bus network.

- Unsymmetrical Faults (L-G, L-L-G and L-L)
- Symmetrical Faults (L-L-L-G, L-L-L)
- Line Switching
- Transformer Energizing

The FFT plot for L-G fault on bus 1 and transformer energizing is as shown in Fig. 3.8 and Fig. 3.9 respectively.

![FFT plot for L-G fault on bus 1 and transformer energizing](image)

**IV Discussion on results**

**4.1 RMS method**

In this method, the parameters namely minimum rms value, maximum rms value, average rms value and rms magnitude unbalance ratio can be used for the extraction of distinguishable features for the events under consideration. These parameters alone or with combination with other one or more parameters can completely characterize the voltage disturbances.

From Table 3.1, it is evident that minimum rms value of phase A voltage for L-G fault is zero whereas it is more than zero for LLG fault. Similarly, this parameter has different values for different type of events. It is seen that rms magnitude unbalance ratio for transformer energizing at bus 4 is different from all other events and is easily distinguishable.

**4.2 Filter Banks Method**

Fig. 4.1 shows the magnitude of phase A, B and C voltages at bus 1 in the frequency scales 1-5 under various events considered. For phase A voltage, in scales 1, 2 and 3 the features of the events considered are aggregating in two different clusters in space where one cluster constitute feature of an individual event i.e. AB fault and therefore it is easily distinguishable from rest of the events. The second cluster is formed by the features of the events except AB fault. Therefore, the features of the events within the second cluster are not distinguishable from each other.

In scale 4, the events aggregate in three different clusters. The features of the events within each cluster are not distinguishable from each other but can be distinguished from cluster to cluster. In scale 5, the features of the events form two clusters. It is seen that AB fault features remain distinguishable from all other events features’ in all the scales.
ABC-G fault features can be distinguished from the features of the rest of the events. In scales 3, 4 and 5 the features of the events except ABC fault are aggregating in a single cluster.

### 4.3 FFT method

A second harmonic component is observed upon transformer energizing at bus 4. From this information, it is concluded that the FFT method is only helpful where dominant harmonics are present in the network. In the case of energizing the transformer second harmonic component is present in the network and FFT peaks are observed near second harmonic of the power fundamental frequency i.e. approx. 100 Hz.

### VI Conclusion

In this paper an effort has been made to extract the features from voltage disturbance waveforms observed under the influence of different events. The methods which were discussed for feature extraction are namely Root mean square (RMS), Filter Banks and Fast Fourier Transforms. According to the observations discussed about the extracted features from different events, FFT method is only able to distinguish transformer energization from other events. The method of filter banks come out be very effective method in distinguishing the various events from each other. Although in some cases the features of various events are aggregating in the same clusters. And in some cases some events are changing clusters while traveling from one scale to another. Moreover, in this method, the information in the scales 1, 2 and 3 is useful in most of the events. In few cases the information in other scales is useful. Features extracted from the bus where fault was simulated, are more distinguishable. In RMS method, RMS unbalance ratio at all buses under various considered events is useful in distinguishing the features of some of the events. Moreover, minimum, maximum and average RMS values are also useful for the extraction of features. Features are more distinguishable from the observed parameter (voltage waveforms) at the faulted bus frequency i.e. 100 Hz.

### References


How to secure e-mail and instant messages
Bari Siddique and Rozina Siddique
University of Texas at Brownsville, TX
and
Mohammad Amin
National University, San Diego, CA

Abstract:
Electronic mail (e-mail) and Instant Messaging (IM) have become indispensable tools in business management and personal relations. It is becoming increasingly vulnerable to attacks such as DoS (Denial of Service), DHA (Directory Harvest Attack), Spam, Phishing, Spyware, Web bug, and Cookie Exploit. As a result, the risk of them causing harm to the security of e-mail and IM is greater than ever. Internet technology is bringing different techniques and devices to manage the threat, however, they are becoming progressively more sophisticated and that is creating a big headache for the IT managers. This paper addresses some of the threats of the e-mail and instant messaging, how they work, and how these threats can be minimized.

Keywords
Chat, Cookie Exploits, Digital Certificate, DoS, DHA, Electronic mail, E-mail client software, Encryption, IMAP, Flash Cookies, Instant Messaging (IM), Internet Technology, Keylogger, Mail Server, Malware, MIME, Phishing, POP, SMTP, Security of Internet, Spam, Spyware, Web bug

Introduction:
Electronic mail and instant messaging have become indispensable tools in business management and even in personal relations, all but replacing traditional means of communication. But as with any widely used tools, these are becoming increasingly vulnerable to malware (spyware, hoaxes, phishing, spam, etc.) attacks. Internet Technology (IT) managers are working hard to combat the increasing volume of attacks on their enterprises. These attacks are also becoming progressively more sophisticated. As a result, the risks of them causing damage to the business and personal level are greater than ever. The majority of the threats that reach to a business do so through the mail server. There are several reasons for this trend:

- An e-mail is easy to access and manipulate
- The SMTP mail protocol is simple and can be emulated by any Internet user
- Firewall-type corporate security devices do not filter SMTP traffic which reaches e-mail servers
- The mail service is a channel for mass infection, via worms and Trojans that replicate in each target, using infected computers and reading mail lists in the host computer.

How do E-mail & Instant Messaging Works:
Electronic mail (e-mail) is a form of communication of messages, which travel across the internet in small packets and deliver to the recipients, whose addresses are being attached to the corresponding messages. When an e-mail is sent through the internet, it is sent to a mail server, which is a dedicated computer with special software that sorts, stores, and routes mails. E-mail is based on store-and-forward technology—a communications method in which data that cannot be sent directly to its destination is temporarily stored until transmission is possible. It determines from the recipient’s address one of the several electronic routes on which to send the message. The message is routed from one system to another and is passed through several mail servers. Each mail server determines the next leg of the message’s voyage until it reached to the final destination—recipient’s mailbox. The following e-mail protocols are used today:

- POP (Post Office Protocol) handles incoming messages. It temporarily stores new messages on a mail server. When a user connects to his/her ISP and requests mail, it is downloaded from the mail server and stored on the user’s computer. Using POP
requires e-mail client software, such as Microsoft Outlook, Mozilla Thunderbird (open source), or QUALCOMM Eudora. The client software provides an inbox for holding incoming message and an outbox for temporarily holding outgoing message before transmission.

- **IMAP (Internet Messaging Access Protocol)** is similar to POP, except we have the option of downloading mail or leaving it on the server.
- **SMTP (Simple Mail Transfer Protocol)** decides which paths an e-mail message takes on the internet, and it handles outgoing messages. It routes outgoing mails to the next mail server or to the destination address.
- **Web based e-mail** keeps mail at a Web server, where it can be accessed using a Web browser. It provides better mail security, and can be accessed from any computer. However, since mails store in a server, we have less control over who access the mails.
- **MIME (Multipurpose Internet Mail Extensions)** specifies how to encode nontext data. It provides a way of disguising digital photos, sounds, and other media as plain ASCII code which can travel over the internet as e-mail attachments. An electronic message incorporated in the e-mail header provides e-mail software with information that allows it to reconstruct the attachment into its original form.

**Instant Messaging (IM)** – Unlike e-mail, **Instant messaging** (IM) is a real-time communication which allows users to exchange short messages over the internet. The instant messaging service will alert a user if somebody on the user's list of correspondents is on-line. IM messages are exchanged directly almost instantly, allowing for a two-way communication in real-time. IM software such as **AOL Instant Messenger**, **Yahoo! Messenger**, **Google Talk**, **Apple iChat**, and **Windows Live Messenger** are regularly used to exchange messages, pictures, videos, sound, data, and programs between friends, families, and coworkers. Messages are typed into Instant Messaging Software, which uses messaging protocols such as **IRC** (Internet Relay Chat) to break the message into packets and ship them to the server for distribution.

**Security of E-mail and Instant Messaging (IM)**

Security of e-mail and IM depends mainly on security for client’s computers, communication channel security, and security for server computers [1]. **Encryption** techniques can protect data transmission, and **digital certificate** can authenticate sender’s and receiver’s identity. **Secure Socket Layer (SSL)**, a widely used protocol, can establish secure internet connection and firewall systems can protect server and client’s computers [1].

**E-mail Security threats:** E-mail traffic is based on the SMTP protocol, which offers little or no reliable safeguards when it comes to exchanging information over the internet between two nodes. The following sections will examine some of the attacks and its possible prevention.

- **Denial of Service (DoS) Attacks:** An attack on a mail server can involve massive sending of connection requests to the server. This means that large communication volumes are generated (frequently from different sources) without even an e-mail being sent. The objective of DoS attacks is to slow down the e-mail server, and, if possible, render it inoperative with the corresponding financial consequences.
- **Directory Harvest Attacks (DHA):** DHA is a technique used by hackers to capture the mail directories of the targeted organization. They do this with software that generates random e-mail addresses, using feasible combinations (common names, positions, department names, etc.). By mass-mailing to these types of addresses and using a trial and error technique, hackers can capture not just e-mail addresses, but also sensitive information such as organizational structure and drives with restricted information.
- **Phishing:** This term describes how malicious users pass themselves off as someone else (normally a company) in order to obtain confidential information from the recipient of the ‘phishing e-mail’. Typically phishers send e-mails that appear to come from a bank or
financial institution and under some pretext or other, ask the recipient for confidential information, such as account access codes. Spoofing of the third-party Web page (which victims are led to through a link in the e-mail) is sometimes highly accurate. This deceitful practice has a high level of success, with often devastating financial consequences.

- **Spyware**: Spyware is computer software that is installed surreptitiously on a personal computer to intercept or take partial control over the user's interaction with the computer, without the user's informed consent. Spyware provides another form of obtaining confidential information from the recipient of e-mail. Once installed, it monitors activities of the installed computer and sends a summary back to third parties. Spyware is also capable of monitoring keystrokes for password, SSN number, and credit card information. *Antispyware* such as Spy Sweeper, Pest Control, Microsoft Antispyware etc. can neutralize this threat.

- **Cookie Exploits**: Marketers, hackers, and pranksters can track the activities of a system using *ad-service cookies* which are created by a third party and activated upon clicking the ad. This can be prevented by blocking the cookies through browser’s security settings. But blocking cookies may prevent doing online purchasing, participating online training classes, using web based mail, or registering for premium services at search engine sites. Blocking selective sites and periodically deleting cookies from the hard disk reduces the exploitation. When conventional cookies are blocked *Flash cookies* are used to track the activities. These cookies can be deleted from the system after locating the #Shared Objects file in which they are stored. *Antispyware* such as Spy Sweeper, Pest Control, Microsoft Antispyware etc. can neutralize ad-serving cookies.

- **Web Bug**: A clear almost invisible GIF 1x1 pixel graphic embedded in e-mail message is designed to track who’s reading the e-mail. Junk e-mail messages use Web bugs for monitoring the number of people who view the e-mail and use that data to determine the direction of marketing campaigns. Once visited a Web bug infested page, an HTTP set-cookie request automatically goes to a server controlled by the marketer or hacker. Web bug detector like Bugnosis can work with the browser to test all the graphics of the page and clear them. These detectors are often included in *antispyware* such as Spy Sweeper, Ad-Aware, Microsoft Antispyware etc.

- **Spam**: Spam, unwanted electronic junk mail, contains Web bugs, viruses, worms, or keyloggers that can wreak havoc on a computer system or steal personal information such as password. It can also be used for phishing scams. A spam filter, utility software, can filters out unwanted e-mails by checking e-mail headers and messages. Security Suites software like Norton Internet Security and McAfee Security Center come with spam filter and a set of basic rule it uses to identify spam. These rules can be updated by downloads, and also creating own set of rules to target spam.

For security of personal computers, Norton Internet Security by Norton, and McAfee Security Center by McAfee can prevent most of the attacks discussed here. For enterprise mail server security, technologies like spam firewall & Web based firewall by Barracuda, Ninja: Total E-mail security by Sunbelt software[2], E-mail security by Sonicwall [3], and Managed services by Panda [4] can block most of these threats.

**Instant Messaging Security**: IM are an increasingly common channel for the spread of malware such as viruses, worms and spyware. While IM enjoys tremendous popularity by enabling real-time communications between friends & families, co-workers, and business partners, it also brings significant risks. These risks fall into three major areas:

- **Inbound threats**: IM creates new vectors for the distribution of malware (viruses, worms, spyware, rootkits, and more) and SpIM (Spam over IM) which can cause a major drain on resources.

- **Outbound threats**: IM opens new 'holes' through which information can leak or be leaked, leading to user privacy concerns and the potential loss of intellectual property.

- IM creates invisible communications channels that operate below the radar of conventional information security measures, exposing the system to regulatory compliance breaches.

- IM clients use port crawling - the ability to exploit any open port on the firewall - so blocking the 'usual' port for the particular application doesn't work.
One of the increasingly popular ways that viruses spread is through instant messaging services such as AIM, Yahoo, and MSN. Because these viruses often spoof the identity of friends on a buddy list, users of these services are highly susceptible to infection unless proper precautions are taken. Here are several rules for using instant messaging services securely:

- On IM, it is very difficult to know beyond a shadow of a doubt that you are actually talking to the person you think you are. Therefore, look at every link and every file transfer with great suspicion.
- Keep your software up to date. However, process of updating can create an opening for a virus infection or compromise. Make sure that you are using the most recent version of the software.
- Do not expect your IM conversations to be private. Instant messaging, like email, is not a secure and private means of sharing information. Messages on an instant messaging service are not encrypted and also must pass through another server before they reach the person you are talking to. This setup makes your instant messages an easy target for someone to monitor.

**Enterprise E-mail and IM protection:** Security protection of e-mail and IM is vital in business community. Enterprise prevention technologies are evolving rapidly to tackle the challenge, some of those are listed below:

- Barracuda networks (www.barracudanetworks.com) provides the following hardware/software solution:
  - Barracuda Spam Firewall for controlling spam.
  - Barracuda IM Firewall for protecting inbound and outbound messages.
  - Barracuda Web application Firewall for Web Site based e-mail and IM protection.
- Sunbelt Software (www.sunbelt-software.com) provides Ninaja E-mail Security which is a total e-mail security solution.
- Jabber (www.jabber.com) brings following IM solution:
  - Jabber XCP for Instant Messaging
  - Jabber Now for Enterprise IM
  - Jabber Clients for multi user chat, desktop, and Web Clients
- Alpha Media (www.alphamedia.net) introduces low end IM solution called PinkNotes Plus. It can be downloaded for 30 days free of charge and it is very cost effective.

**Guidelines for security:**

- Use antispyware utility software to clean up any spyware and run it in the computer in a regular interval.
- Set the browser to reject third-party cookies and register to reject Flash cookies.
- Install an antispoofer tool to identify fake Web sites.
- Set up a disposable e-mail address and use it as necessary.
- Do not click links in untrusted e-mail or pop-up ads, and never respond to e-mail offers, especially those that seem too good to be true.

**References**

Database Panel Discussions: Database Systems for Health Informatics

Moderator: Jacqueline Caesar

Panelists: Mudasser Wyne, Radha Nandkumar, Oswald Crasta, Maryam Davodi Far and Arun Datta

This panel will discuss some critical issues of health informatics including how the patient database can be used efficiently in order to manage, organize, share and secure healthcare related data.
Integrating Multiple Data Sources Into One Analytical Environment

Imad Birouty
Teradata Corporation

Summary:
Integrating multiple subject areas into one data repository brings incredible value in the types of new business questions that can be addressed. This talk will provide an introductory discussion into the new business value that can be realized when integrating multiple data sources into one analytical environment.
Abstract - One of the most important problems in robot kinematics and control is, finding the solution of Inverse Kinematics. As the complexity of robot increases, obtaining the inverse kinematics is difficult and computationally expensive. In this paper, using the ability of ANFIS (Adaptive Neuro-Fuzzy Inference System) to learn from training data, it is possible to create ANFIS with limited mathematical representation of the system. Computer simulations conducted on 2DOF and 3DOF robot manipulator shows the effectiveness of the approach.

Index Terms-- ANFIS, manipulator, Inverse kinematics, Degree of freedom (DOF)

I. Introduction

Robot control actions are executed in the joint coordinates while robot motions are specified in the Cartesian coordinates. Conversion of the position and orientation of a robot manipulator end-effector from Cartesian space to joint space, called as inverse kinematics problem, which is of fundamental importance in calculating desired joint angles for robot manipulator design and control. For a manipulator with n- degree of freedom, at any instant of time joint variables is denoted by \( \theta_i(t) \), \( i = 1, 2, 3, \ldots, n \) and position variables by \( x_j = x(t), \ j = 1, 2, 3, \ldots, m \). The relations between the end-effector position \( x(t) \) and joint angle \( \theta(t) \) can be represented by a single solution depending on the starting point and will not work near singularities. If the joints of the manipulator are more complex, the inverse kinematics solution by using these traditional methods is a time consuming.

Utilization of Neural network (NN) and Fuzzy logic for solving the inverse kinematics is much reported [8-14]. In this paper, neuro-fuzzy systems which provide fuzzy systems with automatic tuning using Neural network is used to solve the inverse kinematics problem. The paper is organized as follows, in section 2, the structure of ANFIS used is presented. Section 3 describes results and discussion. Section 4 ends with conclusion.

II. ANFIS Architecture

This section introduces the basics of ANFIS network architecture and its hybrid learning rule. Adaptive Neuro-Fuzzy Inference System is a feedforward adaptive neural network which implies a fuzzy inference system through its structure and neurons. Jang was one of the first to introduce ANFIS [15]. He reported that the ANFIS architecture can be employed to model nonlinear functions, identify nonlinear components on-line in a control system, and predict a chaotic time series. It is a hybrid neuro-fuzzy technique that brings learning capabilities of neural networks to fuzzy inference systems. The learning algorithm tunes the membership functions of a Sugeno-type Fuzzy Inference System using the training input-output data. A detailed coverage of ANFIS can be found in [2, 37, 14].

For a first order Sugeno type of rule base with two inputs \( x, y \) and one output, the structure of ANFIS is shown in figure (1). The typical rule set can be expressed as,

Rule 1: If \( (x_i \in A_i) \) AND \( (x_j \in B_j) \),
THEN \( f_1 = p_1 x + q_1 y + r_1 \) \tag{3}

Rule 2: If \( (x_i \in A_i) \) AND \( (x_j \in B_j) \),
THEN \( f_2 = p_2 x + q_2 y + r_2 \) \tag{4}

In the first layer, each node denotes the membership functions of fuzzy sets \( A_i, B_i, i=1,2, \) be, \( \mu_{Ai}(x_i), \mu_{Bj}(x_j) \).

In the second layer the T-norm operation will be done related to AND operator of fuzzy rules. Considering T-norm multiplication:

\[
\mu = \min(\mu_{Ai}, \mu_{Bj})
\]

In the third layer, the average is calculated based on weights taken from fuzzy rules,

\[

\bar{\mu} = \frac{\mu}{\mu_{A1} + \mu_{B1}}
\]
Initialize the fuzzy system
Use \texttt{genfis1} or \texttt{genfis2} commands

Give the parameters for learning
Number of Iterations (epochs)
Tolerance (error)

Start learning process
Use command \texttt{anfis}
Stop when tolerance is achieved

Validate
With independent data

III. Simulation and Results

Figure 3(a) and 3(b) shows the two degree of freedom (DOF) and three DOF planar manipulator arm which is simulated in this work.

A. Two Degree of freedom planar manipulator

For a 2 DOF planar manipulator having \( l_1 \) and \( l_2 \) as their link lengths and \( \theta_1, \theta_2 \) as joint angles with \( x, y \) as task coordinates the forward kinematic equations are,

\[ x = l_1 \cos(\theta_1) + l_2 \cos(\theta_1 + \theta_2), \]  

\[ y = l_1 \sin(\theta_1) + l_2 \sin(\theta_1 + \theta_2) \]  

and the inverse kinematics equations are,

\[ \theta_1 = \arctan2(y, x) - \arctan2(k_2, k_1) \]  

\[ \theta_2 = \arctan2(\sin \theta_2, \cos \theta_2) \]

where, \( k_1 = l_1 + l_2 \cos \theta_2, k_2 = l_2 \sin \theta_2 \)

\[ \cos \theta_2 = (x^2 + y^2 - l_1^2 - l_2^2) / (2l_1 l_2) \]  

\[ \sin \theta_2 = \pm(1 - \cos^2 \theta_2)^{1/2} \]
Fig.3 (a) 2 DOF manipulator and (b) 3 DOF manipulator.

Considering length of first arm $l_1 = 10$ and length of second arm $l_2 = 7$ along with joint angle constraints $0 < \theta_1 < \pi/2$, $0 < \theta_2 < \pi$, the x and y coordinates of the arm are calculated for two joints using forward kinematics. The coordinates and the angles are used as training data to train ANFIS network with Gaussian membership function with hybrid learning algorithm. Figure 4 shows the training data of two ANFIS networks for two joint angles.

The coordinates act as input to the ANFIS and the angles act as the output. The learning algorithm "teaches" the ANFIS to map the co-ordinates to the angles through a process called training. In the training phase, the membership functions and the weights will be adjusted such that the required minimum error is satisfied or if the number of epochs reached. At the end of training, the trained ANFIS network would have learned the input-output map and it is tested with the deduced inverse kinematics. Figure 5 shows the difference in theta deduced and the data predicted with ANFIS.

B. Three Degree of freedom planar manipulator

For a 3 DOF planar redundant manipulator, the forward kinematic equations are,

$$x = l_1 \cos(\theta_1) + l_2 \cos(\theta_1 + \theta_2) + l_3 \cos(\theta_1 + \theta_2 + \theta_3)$$  \hspace{1cm} (13)

$$y = l_1 \sin(\theta_1) + l_2 \sin(\theta_1 + \theta_2) + l_3 \sin(\theta_1 + \theta_2 + \theta_3)$$  \hspace{1cm} (14)

$$\phi = \theta_1 + \theta_2 + \theta_3$$  \hspace{1cm} (15)
and the inverse kinematics equations are,

\[ \theta_2 = \text{atan2}(\sin \theta_2, \cos \theta_2) \]  
\[ \theta_1 = \text{atan2}((k_1 y_n - k_2 x_n), k_1 x_n - k_2 y_n) \]  
\[ \theta_3 = \phi - \theta_1 + \theta_2 \]  

where, \( k_1 = l_1 + l_2 \cos \theta_2, k_2 = l_2 \sin \theta_2 \)

\[ \cos \theta_2 = \frac{x^2 + y^2 - l_1^2 - l_2^2}{2l_1 l_2} \]

and \( x_n = x - l_3 \cos \phi \) and \( y_n = y - l_3 \sin \phi \)

For simulation, the length for three links are \( l_1 = 10, l_2 = 7 \) and \( l_3 = 5 \) with joint angle constraints \( 0 < \theta_1 < \frac{\pi}{3}, 0 < \theta_2 < \frac{\pi}{2}, 0 < \theta_3 < \pi \), the same procedure is repeated. Figure 6 shows the training data of three ANFIS networks for three joint angles. Figure 7 shows the difference in theta deduced and the data predicted with ANFIS.

Fig. 6 Training data of (a) \( \theta_1 \) (b) \( \theta_2 \) (c) \( \theta_3 \).

Fig. 7 Difference in theta deduced and the data predicted with ANFIS trained.

IV. Conclusion

The proposed method results in an acceptable error. Trained ANFIS can be utilized to provide fast and acceptable solutions of the inverse kinematics thereby making ANFIS as an alternate approach to map the inverse kinematic solutions. Other techniques like input selection and alternate ways to model the problem may be explored for reducing the error further.

References


Intelligent Cash Management System for an ATM network

Rimvydas Simutis, Kaunas University of Technology, Lithuania, Rimvydas.Simutis@ktu.lt
Darius Dilijonas, Vilnius University, Lithuania, Darius.Dilijonas@vukhf.lt
Lidija Bastina, JSC Penki kontinentų bankinės technologijos, Lithuania, LidijaB@5ci.lt

Abstract. A contribution of this paper lays in novel application of artificial neural network (ANN) based models for cash management optimization for an automatic teller machines (ATM) network. Initially the artificial neural networks are designed for every ATM in network to forecast the money demand for chosen time interval (days, weeks). The ANNs then are retuned periodically using the last observations from ATM reports. The generalization properties of the ANN are improved using special regularization term, which is adapted online depending on complexity of relationship between input and output variables. Performed simulation studies and experimental tests showed good forecasting capacities of the proposed method. The forecasting results are used to optimize the cash upload for every ATM in network. At the current stage the proposed procedure is in the implementing phase for intelligent cash management of ATM network in Lithuania.

Keywords: Automatic Teller Machines, Intelligent Cash Management, Neural Networks

Introduction

Automatic teller machines are computerized telecommunication device which provide a financial institution's customers a method of financial transactions in a public space without the need for a human clerk. According the estimates developed by ATMIA (ATM Industry Association) the number of ATMs worldwide in 2007 was over 1.6 million. The larger the ATM and branch network, the more important becomes proper currency management ensuring that no excess cash is circulating in the network. Serving the ATMs network is a costly task: it takes employees’ time to supervise the network and make decisions about cash management and it involves high operating costs (financial, transport, handling, insurance etc.) [1, 2].

As interest rate rises and greater operating efficiencies become paramount, more banks are turning their attention to driving greater efficiency in how they manage their cash at ATMs [3, 4]. Some banks typically maintain as much as 40% more cash at their ATMs than what's needed, even though many experts consider cash excess of 15% to 20% to be sufficient. Cash-related costs represent about 35-60 % of the overall costs of running an ATM. Through currency
management optimization, banks can avoid falling into the trap of maintaining too much cash and begin to profit by mobilizing idle cash.

Effective currency management and control starts with an intelligent solution that uses advanced algorithms to accurately predict currency supply and demand, allowing banks to forecast demand and proactively manage currency throughout their network. An intelligent cash management system provides the bank the opportunities to lower its operational expenses and improve the return on its cash assets. In Lithuania, the ATM networks are expanding strongly last time, as a result the development of intelligent systems for monitoring and optimization of ATM networks becomes very relevant.

By the end of 2006 the JSC „Penki Kontinentų Bankinės Technologijos“, Lithuania received financial support from the EU structural funds for development of intelligent cash management and optimization system for an ATM network (ASOMIS project). In this paper we present some results by solving this task.

This paper is structured as follows. In section 1, problem formulation and existing approaches are discussed. In section 2, cash demand forecasting model based on flexible artificial neural network is presented. In section 3, the paper describes the cash upload optimization procedure. In section 4, some experimental results from simulation runs are presented and analyzed. Finally, the main results and conclusions of this work are discussed in section.

1. Existing approaches for ATM cash management

The basic element in development of efficient ATM cash management system is a cash demand forecasting model for every ATM. Generally this forecasting model can be created based on historical cash demand data. The historical cash demand for every ATM varies with time and is often overlaid with non stationary behavior of users and also with additional factors, such as paydays, holidays, and seasonal demand in a specific area. Cash drawings are subject to trends and generally follow weekly, monthly and annual cycles. For example, people tend to draw relatively large sums of cash at the beginning of each month. Before Christmas, drawing rates soar, whereas in August, during the summer holidays, rates tend to drop considerably. ATMs that are located in shopping centers, for example, are most heaped on Fridays and Saturdays. Consequently, the development of the forecasting model is complicated procedure, because it must consider the changing behavior of users and various input information.

Based on the cash demand forecasting model the optimization procedure determines the optimum cash amount for each ATM by calculating the transport and money upload costs against interest rates. Cash management system has to guarantee the availability of cash in the ATMs network, should estimate optimal amount of stocked money plus efficiently manage and control day-to-day cash handling, also money transportation with reducing of currency transportation and servicing costs. The system should be flexible enough to allow the bank to re-forecast future demand, perform WHAT – IF analyses, and optimize the network as the cash distribution environment evolves.

Most known cash management systems for ATM network are presented in Table 1. The solutions presented in this table have the following drawbacks:
- Cash demand forecast for every ATM generally is based on linear regression models with seasonality coefficients. The development of such models is relatively complicated and differs for various ATM. Therefore preparation of forecasting models for whole ATM network is difficult task for owners of machines;
The parameters of forecasting models are determined in the system implementation stage and are hold constant during the operation phase. On the other hand, business environment changes continually in a real world and, consequently, the model parameters must be also adapted to the changing environment. To eliminate these weaknesses, we propose a new forecasting method based on flexible artificial neural networks. The functioning principles of these methods are discussed below.

### Table 1. Most known solutions for cash management of ATMs’ network

<table>
<thead>
<tr>
<th>Company</th>
<th>Software Product</th>
<th>WWW page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transoft International</td>
<td>OptiCa$h</td>
<td><a href="http://www.transoftinc.com/site/index.php?option=com_content&amp;task=view&amp;id=18&amp;Itemid=40">http://www.transoftinc.com/site/index.php?option=com_content&amp;task=view&amp;id=18&amp;Itemid=40</a></td>
</tr>
<tr>
<td>Wincor Nixdorf</td>
<td>Pro Cash Analyser</td>
<td><a href="http://www.wincor-nixdorf.com/internet/com/Products/Software/Banking/CashManagement/Main.html">http://www.wincor-nixdorf.com/internet/com/Products/Software/Banking/CashManagement/Main.html</a></td>
</tr>
</tbody>
</table>

### 2. Flexible neural network for cash demand forecasting

Artificial Neural Networks (ANNs) are universal and highly flexible function approximators first used in the fields of cognitive science and engineering. Furthermore they are used for tasks such as pattern recognition, classification and time series forecasting [5,6,7]. In recent years, ANN becomes increasingly popular in financial markets. The key to all forecasting applications is to capture and process the historical data so that they provide insight into the future. Cash demand forecasting is main element for creation of efficient ATM/branch network cash management and optimization systems. The new advanced approaches for cash forecasting are fuzzy expert systems and artificial neural networks. In this paper, we are concentrating on application of artificial neural networks for cash forecasting problem. The general idea behind the use of ANN in cash forecasting is to allow the network to map the nonlinear relationships between various factors affecting the cash withdrawal and the actual cash demand. Once this relationship between inputs and outputs is identified, it gives the cash forecast using values for various factors affecting the cash withdrawal as ANN inputs.

One of the most important components in the success of neural network solution is the structure of the ANN and the data necessary to train the network. In this study, we used simulated data and real data for training and evaluation of the artificial neural networks. For every ATM machine a separate three-layer feed-forward neural network was designed. The neural network was trained using Levenberg-Marquardt optimization method and RMS (root mean square) error between predicted and real value. Regularization term was also included in the training criterion [5,6]. The input variables for ANN were coded values of weekday, day of the month, month of the year, holiday effect value and average daily cash demand for ATM in last week. The output variable of ANN was cash demand for the ATM for the next basic time interval. Two types of neural network were prepared for each ATM: an ANN with basic forecasting time interval - one day, and an ANN with basic forecasting time interval - one week. To forecast the cash demand
for the following time intervals, the predicted values from past time interval were supplied in recurrent mode for the forecasting model. One day basic time interval was used for short time prediction of ATM’s cash demand (typically 3-7 days), and one week basic time interval was used for long time prediction (typically 2-6 weeks). For simplification purpose the ANN structure for all ATM in the network was chosen the same (the same inputs and the same number of hidden units in ANN). The number of neurons in the hidden layer was chosen relative big (15 hyperbolic tangents neurons in hidden layer). Such neural network can approximate very complicated relationships between input output variables but the generalization properties of neural network can be very pure. Therefore we proposed a special flexible neural network design procedure for cash demand prediction for every local ATM. The realization of the proposed procedure is executed in the following steps:

1) Assemble input-output data from every local ATM (historical data from 2-3 years is usually necessary for reliable training of ANN);
2) Divide assembled data in training (70%) and testing sets (30%);
3) Train ANN using Levenberg - Marquard optimization method and choose various values of regularization term, which is included in the training criterion;
4) Estimate the normalized sum square error (NSSE) of ANN for test data set;
5) From all training runs choose the regularization term which gives the minimum of NSSE in test data, and use it as optimal regularization term;
6) Repeat ANN training on whole data sets using optimal regularization term and use this ANN as basis for cash demand prediction;
7) As the additional portion of new data about functioning of local ATM is available (typically in one week), repeat the steps 2-6 with less number of training iteration;
8) Use the fresh adapted ANN for cash demand prediction in chosen time interval.

The proposed algorithm adapts the ANN parameters (weights) following the new observation; therefore the designed ANN is always tuned to the current situation observed in the business environment. Having the models to forecast the daily (or weekly) cash demand for every ATM, it is possible to monitor, plan and optimize the cash uploads for every ATM in network.

3. Optimization of ATM’s cash uploads

The proposed ANN-based algorithm forecasts the cash demand during the chosen time interval for each ATM on an individual basis. Using this information the optimization procedure for ATM cash uploads can be implemented. The proposed cash upload optimization procedure for an ATM contains the following steps:
- The amount of cash positions on every ATM is monitored daily;
- Based on the trained ANN a cash demand for \( n \)-subsequent day (weeks) is forecasted for every ATM;
- If the cash position in ATM for next day (week) is smaller as required, the optimization algorithm for cash upload is activated;
- Using an optimization algorithm, the optimal cash upload for ATM is estimated. The algorithm searches for minimal ATM’s maintenance cost function. This cost function is sum of cash costs (annual interest rate), cash uploading costs and constant ATM-service costs. We used simulated annealing optimization method [5] to estimate an amount of cash upload for ATM which minimizes the ATM’s maintenance cost function.
In this way optimization procedure determines the lowest cost of cash distribution for every ATM in network. It is based on flexible forecasting model for every ATM and provides the opportunity for a bank to lower its operational expenses and improve the return on its cash assets.

4. Simulation studies and experimental tests

To test the possibilities of artificial neural network to forecast the cash demand for ATM, a simulation environment for ATM was designed. A behavior of typical ATM was simulated using weekly and monthly seasonality along with long term trends and special events (holiday effects). The simulation environment has imitated the money withdrawal from typical ATM in Kaunas city, Lithuania. Simulation of ATM and training of ANN was realized in MATLAB programming environment, using NNSYSID neural network toolbox [7]. Typical simulation results for one ATM are presented in the figure 1.

![Figure 1. Typical example of daily cash demand for local ATM](image)

Initial neural network was trained using data records from 2 years. After that, ANN was used for daily prediction of cash demand for ATM. Every week ANN was also retrained using moving window data from the last two years. Proposed ANN training procedure allowed preparing an ANN, which was able to predict the cash demand for the next day with high accuracy. Mean average proportional error (MAPE) of daily cash demand prediction for various simulation runs varied between 1.5 - 2 %. Typical example of the prediction results is presented in figure 2.

The proposed procedure was also tested using real cash demand data from ATM network. In this case the cash demand prediction error was significantly higher. MAPE for daily prediction
for various ATM fluctuated between 15-18%, MAPE for weekly prediction was in range 8-15%. Typical daily prediction results are presented in figure 3.

![Figure 2. Illustration of cash demand prediction quality using simulated data (*- simulated data, o-prediction)](image)

Figure 2. Illustration of cash demand prediction quality using simulated data (*- simulated data, o-prediction)

![Figure 3. Illustration of daily cash demand prediction quality using data from real ATM (*- real data, o-prediction)](image)

Figure 3. Illustration of daily cash demand prediction quality using data from real ATM (*- real data, o-prediction)

The prediction results using flexible ANNs were significantly better when comparing with prediction results using linear models with seasonality coefficients (5-10% prediction errors for simulated data and 25-30% prediction errors for real data). As a result these models can be
used efficiently for monitoring and supervision of ATMs in network and also for ATMs’ cash upload optimization.

To test the possibilities of forecasting models and to evaluate the efficiency of the cash upload optimization procedure, a simulation environment for an ATM network was designed. Then the system of an ATM network, which consists of 1225 ATMs, was simulated using typical cash demand behavior discussed previously in this paper. To evaluate the efficiency of the optimization procedure, two different scenarios with various annual interest rate and cash uploading costs were simulated. The parameters for modeling scenarios are given in Table 2. Table 2. Two modeling scenarios for evaluation the efficiency of cash optimization system (1$ =2.5 Litas)

<table>
<thead>
<tr>
<th>Scenario I</th>
<th>Scenario II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of ATMs = 1225</td>
<td>Number of ATMs = 1225</td>
</tr>
<tr>
<td>Annual interest rate = 6 %</td>
<td>Annual interest rate = 3.5%</td>
</tr>
<tr>
<td>Cost of cash uploading = 150 LT/ATM</td>
<td>Cost of cash uploading = 300 LT/ATM</td>
</tr>
<tr>
<td>Maximal amount of money loaded in ATM = 1 000 000 LT</td>
<td>Maximal amount of money loaded in ATM = 1 000 000 LT</td>
</tr>
<tr>
<td>Average daily cash demand = 200 000 LT/ATM</td>
<td>Average daily cash demand = 200 000 LT/ATM</td>
</tr>
<tr>
<td>Constant maintenance costs 30 LT/ATM/Day</td>
<td>Constant maintenance costs 30 LT/ATM/Day</td>
</tr>
</tbody>
</table>

Simulation of ATMs network and realization of ANN based forecasting methods were implemented using MATLAB programming environment and NNSYSID neural network toolbox [7].

5. Results and conclusions

The proposed forecasting model based on flexible artificial neural network training procedure allowed predicting the cash demand for ATMs with good prediction quality. Mean average proportional error (MAPE) of daily cash demand prediction for various simulation runs varied between 1.5 - 2 % for simulated data and between 15-18% for real data. This accuracy is satisfactory to perform real optimization procedure for ATMs’ cash uploading estimations.

Figure 4 shows the efficiency of the cash optimization procedure for scenario I. The simulation was carried out for two years. During the first year the ATMs in the network were loaded when amount of cash in ATMs achieved minimum restriction. Then the ATMs were loaded with maximum amount of cash. During the second year the proposed optimization procedure was activated and optimal amount of uploading cash was estimated for every ATM machine. Optimization procedure allowed to decrease daily costs for ATM network maintenance approximately 18% (from 161 000 Litas, till 132 000 Litas, Figure 4). During the entire year, it allowed to decrease significantly the ATM network maintenance cost and to save about 10 million Litas.

Figure 5 presents the simulation results for scenario 2. In this case, the cash management optimization is not so efficient and allows decreasing the maintenance costs only by 2 %. As one can see from the simulations, the optimization results depend strongly on money cost (annual interest rate) and cash uploading cost. The proposed system gives very promising results by higher interest rates and lower costs of cash uploading.
Figure 4. Simulation results for cash optimization system (Scenario I): Daily maintenance costs for ATM network before and after optimization (top); b) Total maintenance costs for typical and optimized ATM network (bottom).
Figure 5. Simulation results for cash optimization system (Scenario II): Daily maintenance costs for ATM network before and after optimization (top); b) Total maintenance costs for typical and optimized ATM network (bottom).
Using various simulation experiments together with the proposed optimization procedure, it is easy to test different WHAT-IF scenarios and make the right decision about possible implementation of the cash upload optimization system in real business environment.

At this stage the proposed forecasting model and the cash upload optimization procedure are in the implementing phase for intelligent cash management in real ATM network in Lithuania.

Acknowledgments

This research work has been supported by means of EU structural funds (Project No. BPD2004-ERPF-3.17-06/0045).

References

RIN Evaluation with dispersion compensation for Optimum Performance of Optical Transmission System

Kamaljit Singh Bhatia
Dept. of Electronics and Communication Engineering
DAV institute of Engineering and Technology,
Jalandhar,Punjab,India

Dr. Manoj Kumar
Dept. of Electronics and Communication Engineering
DAV institute of Engineering and Technology,
Jalandhar,Punjab,India

Abstract-- In this paper, we have carried out the simulative analysis for averaging out the optimal value of RIN with dispersion compensation for better performance in high bit rate optical transmission link. The impact of data rate is investigated for optimal performance.

Keywords: Relative Intensity to Noise Ratio, Intensity Noise

I. INTRODUCTION

Today’s optical transmission links are targeting a bit rate in T bits/sec, which is possible only if we are able to control various transmission affecting parameters like attenuation, dispersion, fiber non-linearities and relative intensity to noise (RIN). Focus on development of broadband optical communication system is incredible since it offers combination of wide bandwidth and low losses unmatched by any other transmission medium but besides dispersion, fiber non-linearities and RIN remain inherent limitations to such systems thereby degrading the performance [1-3]. Therefore, in order to realize broadband optical communication systems and networks, it is imperative to study and analyse the impact of RIN. In the present scope of our work, we have focused mainly on RIN. It has been shown that RIN induced by dispersion at first-order dispersion wavelength in single mode fiber transmission lines and small signal frequency response of a linear dispersive single-mode fiber near zero first-order dispersion wavelengths was observed [4-5]. Laser diode modulation and noise, which may be intensity noise or phase noise, limits the system’s performance [6]. A considerable FM-AM noise conversion occurs in dispersive fiber-link, which must be taken into account when designing analogue sub carrier distribution systems [7]. DBF laser RIN degradation occurs in CATV light wave transmission system [8]. For a link-length of 10km at $\lambda$=1.55 $\mu$m, RIN values between -125 and -145 db/Hz are expected for frequencies between 1 and 10 GHz. These RIN values can be reduced either for laser diodes with narrower linewidth or for fiber links with lower dispersion [9]. However, in addition to the FM-AM noise conversion, nonlinear distortion caused by FM-AM conversion must be accounted for in analogue systems. The Influence of fiber nonlinearity on the conversion of laser and optical amplifier phase noise, intensity noise by fiber transmission was investigated in [9]. A very good agreement of RIN spectra at the output of a standard single mode fiber between experimental data and theoretical prediction was achieved. Results reveals that the fiber nonlinearity and dispersion can enhance the RIN magnitude significantly and lead to the shift of RIN dips towards higher frequencies and consequently to a broader RIN spectrum at the fiber output. The theory presented provides a tool in order to reduce RIN in cascaded optical amplifiers systems, by suitable choice of system parameters. It may provide basic information to minimize detrimental RIN effects in future communication systems. Further, the expression for RIN including higher order dispersion term was derived using small signal analysis [10]. It was shown that the second order dispersion term was having negligible effect on RIN. Later on Large signal Analysis of FM-AM conversion in Dispersive optical fibers for PCM systems including Second order dispersion was carried out [11]. However none of the researchers have evaluated RIN for another important parameter laser linewidth with dispersion compensation.

II. THEORY

The output of semiconductor laser exhibit fluctuations in its intensity, phase and frequency even when the laser is biased at constant current with negligible current fluctuations. The two fundamental noise mechanisms are spontaneous emission and electron hole recombination. Noise in semiconductor lasers is dominated by spontaneous emission. Each spontaneously emitted photon adds a small field component to the coherent field (established by stimulation emission), which is random in nature and thus perturbs the both amplitude and phase in a random manner. The occurrence rate of such a spontaneously emitted random field is about $10^{12}$ s$^{-1}$ [1-3]. Because of which intensity and phase of emitted light exhibit fluctuations over a time scale as short as 100ps. Intensity fluctuations lead to the limited signal to noise ratio (SNR) where as phase fluctuations leads to the finite spectral linewidth when semiconductor lasers are operated at constant current. Clearly such fluctuations lead to the degradation of system performance, therefore it is important to estimate their magnitude. Amplitude fluctuations are characterized by a factor called as Relative Intensity to Noise ratio (RIN)

III. SYSTEM DESCRIPTION
The simulation set-up for modeling of RIN determination along with PMD compensation using fiber grating is shown in figure 1.

The transmitter section as shown in the figure 1 consists of data source, electric driver, and laser source and amplitude modulator. The data source is modulated using NRZ data format at different varying data rates. The laser is of the type CW Lorentzian with laser center emission frequency 1550nm (193.4145 THz). The amplitude modulator is of type sine square with excess loss of 3 dB. The output of modulator is fed to optical link consisting of three optical spans of length 50 km each measuring total link length of 150 km having booster, optical splitter and a standard optical fiber. The optical splitter is connected to correlate the signal at input and output of three optical spans using optical spectrum analyzers. The EDFAs are of fixed output power type each with noise figure of 5 dB. The output is detected at the receiver by PIN detector with responsivity 0.875 and is passed through electric filter and output is observed on electroscope. The electric filter is of the type Bessel with -3dB bandwidth equal to 8 GHz. The electroscope gives eye diagram, Q value, bit error rate and eye closure penalty. The effects of fiber non-linearities and polarization mode dispersion are also considered in the simulative analysis.

IV. RESULTS AND DISCUSSIONS

A pseudo random sequence length of bits taken one bit per symbol is used to obtain realistic output values at the receiver. Firstly, to observe the impact of RIN upon system performance, simulation results are obtained for different data rates varying from 2 Gbps to 15 Gbps. It was observed that for the data rate up to 8 Gbps Q value for the system remains independent of RIN. As we increase the data rate impact of RIN upon the Q value, jitter, eye opening etc. come into the picture. It is investigated that system provides best results at data rate of 9.953 Gbps (refer Table 1).

Eye opening is defined as the difference between the minimum values of the samples decided as logical one and maximum value of the sample decided as logical zero. Average eye opening corresponds to difference between the average values for the samples. It is observed that as the linewidth and receiver attenuation is increased, the eye opening decreases. This can be explained on the basis of the fact that increase in linewidth or the receiver attenuation will introduce more dominance to RIN and its cumulative effect with fiber nonlinearities. The ratio of average eye opening to the eye opening expressed in dB is a measure of eye closure penalty. The plot of this penalty and relative intensity to noise parameter (r) is shown in figure 3. Further, RIN is correlated with Q value as shown in figure 4 and it is investigated that its value should be negative.
For positive values of RIN, Q value is found to be very less as compared with the negative values of RIN. In this paper we have iterated the values of RIN from 10 dB/Hz to -180 dB/Hz and different parameters are observed. We found that Q value decreases for negative values of RIN up to around -100 dB/Hz with further decrease in its value Q value increases and again tends to be maximum at -150dB/Hz.

Table 1: Performance indices at different data rates

<table>
<thead>
<tr>
<th>S.No</th>
<th>Data rate/ value</th>
<th>Q value [lin]</th>
<th>BER</th>
<th>Eye Opening</th>
<th>Jitter [ns]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.0</td>
<td>100</td>
<td>1 e-040</td>
<td>0.0066</td>
<td>0.0233</td>
</tr>
<tr>
<td>2.</td>
<td>5.0</td>
<td>100</td>
<td>1 e-040</td>
<td>0.010</td>
<td>0.022</td>
</tr>
<tr>
<td>3.</td>
<td>8.0</td>
<td>4.31</td>
<td>1.169 e-040</td>
<td>0.0008</td>
<td>0.024</td>
</tr>
<tr>
<td>4.</td>
<td>9.953</td>
<td>100</td>
<td>1 e-040</td>
<td>0.008</td>
<td>0.024</td>
</tr>
<tr>
<td>5.</td>
<td>12</td>
<td>3.637</td>
<td>0.00013</td>
<td>0.0009</td>
<td>0.024</td>
</tr>
<tr>
<td>6.</td>
<td>15</td>
<td>3.33</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.024</td>
</tr>
</tbody>
</table>
Figure 4: Eye pattern for RIN = -150 dB/Hz and Data rate = 9.953 Gbps

Figure 5: Response of Average Eye opening w.r.t RIN

Figure 6: Output optical spectrum for RIN = -150 dB/Hz and Data rate = 9.953Gbps

Figure 7: Output Electrical spectrum for RIN = -150 dB/Hz and Data rate = 9.953Gbps
V. CONCLUSION

In this paper, we have concluded that Data rate of pulse has a remarkable effect upon system performance. It is investigated that increase in data rate results in increase in RIN and hence performance of system degrades. RIN values for a link length of 150 km were obtained while taking into account the fiber non-linearities and compensating polarization mode dispersion effect. We investigated the optimal values for data rate and RIN for better performance. The limiting value of data rate should be 9.5Gbps at which performance indices provide best results and RIN value corresponding to this data rate is measured to be -150 dB/Hz. and the average value of RIN is measured to be -120 dB/Hz.

V. REFERENCES


Simulation Study and Implementation of Single Piece Lean Line Design for Cycle Time Reduction - A Practical Approach

Prof. D. Arunkumar  
Asst. Prof, Dept. of Mechanical Engineering,  
S.J. College of Engineering, Mysore. darunkumar63@yahoo.com

Prof. V. Ramesh  
Research Scholar & Asst. Prof, Dept. of Industrial & Production Engg.,  
S.J. College of Engineering, Mysore vajram_ramesh@yahoo.com

Dr. K.V. Sreenivasa Prasad  
Research Guide & Professor & Head, Dept. of Industrial & Production Engg.,  
S.J. College of Engineering, Mysore. kvsprasad3@yahoo.com

Dr. T. R. Srinivas  
Asst. Professor, Dept. of Industrial & Production Engg.,  
S.J. College of Engineering, Mysore. shashisrini@yahoo.com

Abstract
This paper discusses the Simulation study carried out for the reduction of cycle time, number of workers, number of setups and improved layouts using VISIO 2007 & Pro Model- Process simulator software for the assembly unit in a pump manufacturing industry. The study concentrated on introducing the Milk run to ensure the continuous replenishment of raw material at the workstation and KANBAN techniques for the defect free movement of semi finished & finished components during the assembly process. The existing flow line is studied & the alternate flow line is proposed using the Simulation technique. Based on the concepts of Lean Manufacturing; prototype models are fabricated by incorporating the principles of Industrial Engineering techniques. Ergonomics consideration is adopted in the Lean Line Design & Method study is carried out for optimizing the assembly of the pump.


INTRODUCTION

The study discusses the Simulation study carried out in a Pump Manufacturing Industry for optimizing the layout based on the tools of Lean Manufacturing. Lean Simulation in a Manufacturing factory floor is an invaluable tool in the implementation of lean manufacturing. Many manufacturers will not make a change to the process before a simulation is performed to determine the impact of the change. Simulation can be considered as inexpensive insurance against costly mistakes. In this study, a detailed plan of action for the introduction of one-piece Lean Line design for the assembly of a pump unit is discussed in detail. The simulation study is carried out using VISIO 2007 & ProModel- Process Planner software. The study also discusses briefly the KANBAN technique, Milk run concepts used during the implementation of One Piece
Lean Line flow design for the assembly of the Pump Unit. The Implementation of Lean line design for single piece flow line for the assembly of pump has the quantitative aim of increasing productivity & flexibility, lowering the investment ratio, utilizing the space to the optimum extent and optimizing the through-put time.

OBJECTIVE OF THE PAPER

a. To carry out interview with the Training Manager, Project Engineer concerned in the implementation of lean line design.
b. To calculate the number of calibration benches, number of activities, number of tightening units & storage units required for the assembly of the pump.
c. To collect Data using Standard Time study technique for each of the activities.
d. To calculate the number of Supermarket required for the assembly & to discuss the concept of Milk run
e. To carryout simulation analysis using VISIO 2007 software & PROMODEL - PROCESS SIMULATOR software.
f. To carryout Method Study for the assembly of the pump.
g. To carryout Ergonomic study to fabricate the wooden prototype models and to propose Lean Line Design for the implementation in the shop floor for the full scale production.

LITERATURE SURVEY

According to Hank Czarnecki and Nicholas Loyd [1], Manufacturing factory floor simulations are invaluable tools in the implementation of lean manufacturing. Simulation can be considered as inexpensive insurance against costly mistakes; Michelle Eileen Scullin [2] states that, Simulation is used to evaluate the behavioral issues of processes. In a manufacturing realm this means simulation shows how each operation affects other operations so determinations can be made about where bottlenecks or other problems exist in the process. Theoretically, the integration of VSM and Simulation can aide in process improvement by showing both the static and behavioral characteristics of a process. Sanjay Jain & Swee Leong [3], Simulation provides the capability to evaluate performance of a system operating under current or proposed configurations, policies and procedures. Simon Dennis, Ben King, Martin Hind, Stewart Robinson [4] in his paper discusses that using the Simulation model, it is possible to predict what effects the proposed solutions would have on such things as resourcing, quality of service, cost and process efficiency. Yang-Hua Lian [5] in his doctoral research paper states, Simulation models are the final products of atoms, database and model generator. By changing the data in the database, we can yield a simulation model that corresponds to the new data set without much effort. In this way, different scenarios of Value Stream Mapping (VSM) can be transformed into simulation models in a short time and we can easily obtain feedback and implement improvements to the system after analyzing and comparing the outputs of simulation models. The more important features of the simulation model is that the job volumes, resource levels, performance targets, quality of service targets, etc. could be changed before running the model by a non-simulation expert. Diamond, et. al 2002 [6], Identifying and specifying the role of simulation within the lean approach seem valuable and even necessary in expanding the simulation application base. Manar [8] discusses the guidelines for the effective utilization of Simulation tools in the Industrial environment to improve productivity. Adams, et al. (1999)
[7],[9] gives an overview of how simulation could be used within the lean manufacturing strategy, Altinkilic, M. [11] in his paper on “Simulation-based layout planning of a production plant” discusses the parameters to be considered in carrying out layout design & the benefits gained in carrying out simulation analysis before implementing the actual set ups. Carria & Carson [12] have discussed the role of Simulation in Manufacturing. These research papers have been helpful in carrying out simulation analysis & modeling for the Implementation of Lean Line Design Layout for the assembly line for the Pump in a Manufacturing organization in Karnataka State, India.

PROBLEM DEFINITION

At present, the company has got 6 assembly lines in plant 1 for ‘A’ pump. Out of 6 lines, 4 lines are modified with the necessary accessories like quality check devices, torque checking equipments for inspection purposes. The other two lines are 40 years old which needs to be upgraded. In addition, the flow line is outdated & needs to be upgraded to Lean Line Flow line Design. Hence, a Simulation study on the implementation of a flow line for the continuous replenishment of raw material & assembly of the component is carried out.

METHODOLOGY

The study is carried out in a Pump Manufacturing Industry in Bangalore.

a. The decision regarding the number of activities required & the number of calibration benches is calculated mathematically and the logic is based on the number of workers available, space constraint factor, and number of pumps to be manufactured. The target is fixed based on the demand requirements.

b. Using the standard time study charts, the cycle time, change over time is collected

c. Using AUTOCAD 2006, the layout plan for the assembly of pump is drawn. Using the VISIO 2007 software, the models are created based on requirements & also decision entities are suitably incorporated. Using Pro Model Process Planner Simulator software the simulation analysis is carried out to validate the process plan for the assembly of the pump. Suitable care is taken to incorporate Modern Lean Manufacturing Tools. The software permits to input the number of activities involved in the process of assembly, data regarding the cycle time, number of workers, allotment of workers to the appropriate activities, batch size requirements and flow constraints required. Based on the input data, the software carries out the simulation analysis. It displays operation time, idle time in the process, buffer stocks data and displays the results graphically.

d. After alternate simulation studies, the proposed Simulated Layout are fabricated using the wooden card boards and testing is carried out.

CURRENT STATUS

The company is carrying out the assembly of pumps in 6 lines & the layout of the machines are scattered. The number of pumps to be assembled in the existing layout is 2000 per day. The number of hours of working is fixed as 20 hours per day. Hence, Total number of pumps to be assembled is 100 per hour. But, 2 lines out of the 6 lines have become obsolete. There exists lot of space constraint. It was observed that the bottlenecks were high involving wastes in
transportation, underproduction, higher work in process inventories, storage problems. The assembly of PE standard in-line fuel pumps, Distributor fuel injection pumps, axial piston distributor pumps, radial-piston distributors pumps are being carried out in 6 assembly lines. The area occupied by the 1st assembly line unit is around 72 Sq m. The current layout existing for the assembly of the pump is shown in Fig 1.

Fig 1 Existing Layout for the assembly of Pump (1 assembly Line)

a. This line consisted of 21 stations.
b. There was no proper material movement.
c. It was in L shape. This was created in L shape because of space constraint at that time.
d. No torque checking.
The current layout is not having the features of Lean Manufacturing. Hence, the study is concentrated on the implementation of Lean Line Design Layout.

The aim is to create an exclusive one piece flow line with in order to apply the concepts of Lean Manufacturing in pursuit to reduce wastes in the process of manufacturing, the number of stations & number of workers. Further, an attempt is made for the introduction of supermarkets wherever called for. The design concentrated on the introduction of Milk run. A Milk Run concept works on the principle of a water spider continuously visiting the different work centers and supply the materials required for them. At the same time it supplies the finished parts or components to the super market for further processing. This will assist in reducing the Work in Process inventory considerably.

SIMULATION MODEL DEVELOPMENT

The drawing of the layout is drawn after calculating the following parameters.
a) Calculate the number of workbenches required to carry out the activities & categorizing the number of activities required for the assembly.
b) Calculate the number of calibration benches required for the calibration of the pumps assembled.
c) Calculate the number of supermarkets required
d) Provide provision in the layout for the implementation of Milk run for loading & unloading of the component.
e) Decide on the distance between one worktable to another worktable to avoid unnecessary bottlenecks.
a) Proposal I – Integrated Lean Line Design

Preliminary discussion with the Managers and Engineers was carried out & an initial layout was proposed for the assembly of 120 pumps. The proposed layout namely Integrated Lean Line Design Layout is shown in the Figure 2.

The layout proposed is having the following modifications.

- This line reduced the number of stations from the existing 21 stations to 20 stations and also the older L layout was changed to the U layout. The details of the activities carried out are discussed in Table 1.
- The layout introduced the provision for the introduction of Milk Run for the movement of the material. Also, 7 calibration benches, 2 Phasing stations, 2 for tightness checking and 3 for pre packing activities are introduced & the details are shown in Table 2.
- The flow is external U for the assembly activities & internal U for the calibration & testing units. This reduces the space constraint to the maximum extent.
- An exclusive torque checking area is identified.
- A repair station is introduced
- 2 Buffer areas are introduced
<table>
<thead>
<tr>
<th>STATION NO</th>
<th>DESCRIPTION</th>
<th>CYCLE TIME IN MIN (AVERAGE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cleaning the Element &amp; DV</td>
<td>2.53</td>
</tr>
<tr>
<td>2</td>
<td>Assembly of DV holder</td>
<td>2.55</td>
</tr>
<tr>
<td>3</td>
<td>Assembly of Element and DV</td>
<td>2.51</td>
</tr>
<tr>
<td>4</td>
<td>DV holder tightening</td>
<td>1.25</td>
</tr>
<tr>
<td>5</td>
<td>Gallery Tightness Checking</td>
<td>2.83</td>
</tr>
<tr>
<td>6</td>
<td>Assembly of Control Rod and its freeness checking</td>
<td>2.86</td>
</tr>
<tr>
<td>7</td>
<td>Assembly of Roller tappet and lower spring plate and RT clip assembly</td>
<td>3.25</td>
</tr>
<tr>
<td>8</td>
<td>Assembly of inner race of bearing to Gov Housing by pressing and Assembly of gasket on to pump and assembling the Gov Housing to Pump</td>
<td>3.58</td>
</tr>
<tr>
<td>9</td>
<td>Tightening by auto screwing</td>
<td>3.10</td>
</tr>
<tr>
<td>10</td>
<td>Bearing cover flange sub assembly and Assembly of Bearing inner race to cam shaft</td>
<td>2.25</td>
</tr>
<tr>
<td>11</td>
<td>Pressing Cover Bearing</td>
<td>2.58</td>
</tr>
<tr>
<td>12</td>
<td>Assembly of Max screws</td>
<td>2.51</td>
</tr>
<tr>
<td>13</td>
<td>Pressure checking and Height checking</td>
<td>2.25</td>
</tr>
<tr>
<td>14</td>
<td>Cam height over checking</td>
<td>1.10</td>
</tr>
<tr>
<td>15</td>
<td>Check play over cam shaft</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Pressing Base cup</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Clamping Jaw assembly</td>
<td>2.52</td>
</tr>
<tr>
<td>18</td>
<td>Tightening fly weight</td>
<td>1.80</td>
</tr>
<tr>
<td>19</td>
<td>Stud feed pump</td>
<td>2.25</td>
</tr>
<tr>
<td>20</td>
<td>Assembly cover</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td><strong>GRAND TOTAL AVERAGE CYCLE TIME</strong></td>
<td><strong>44.43 Mins.</strong></td>
</tr>
</tbody>
</table>

Table 1 Average Cycle Time taken for 8 Hours daily for 10 Days
<table>
<thead>
<tr>
<th>NO.</th>
<th>TESTING OF THE ASSEMBLED PUMP</th>
<th>CYCLE TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Phasing 1</td>
<td>4 Mins</td>
</tr>
<tr>
<td>22</td>
<td>Phasing 2</td>
<td>4 Mins</td>
</tr>
<tr>
<td>23</td>
<td>Tightness check 1</td>
<td>3 Mins</td>
</tr>
<tr>
<td>24</td>
<td>Tightness Check 2</td>
<td>3 Mins</td>
</tr>
<tr>
<td>25-31</td>
<td>Calibration Checks for 7 Tables</td>
<td>21 Mins</td>
</tr>
<tr>
<td></td>
<td>Note: Avg. Time= 21 Mins/7 tables</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Pre packing activities (2 Mins each for 3 Benches)</td>
<td>6 Mins</td>
</tr>
</tbody>
</table>

**Grand Total Cycle Time**

41 Mins

Table 2 Calibration & testing activities

The factory Layout was drawn using AUTOCAD 2006 software & the drawing was imported to VISIO 2007 software. Using PROMODEL- PROCESS PLANNER software, the cycle time for each activity & the workers are assigned to each activity. A simulation analysis was carried out for 40 hrs in normal run is carried out. Fig 3 shows the Simulation Setup -40 Hours Normal run (VISIO 2007).

![Fig 3 Simulation for 40 hrs Normal runs](image)

The Simulation Analysis was carried out using the PROCESS PLANNER SIMULATION SOFTWARE. The results screen at 20 % zoom for 40 hours of normal run, output diagram & the result data for the Integrated Lean Line design is as shown in Fig 4, 5 & 6 respectively.
Fig 4 Simulation Analysis using ProModel Process Planner Simulator

Fig 5 Output Results of the Integrated Lean Line Layout (Integrated Lean Line Layout)

Fig 6 Sample activities from 1 to 10 (Utilization of the units)
It was observed from the simulation analysis that the layout is congested for the workers easy movement & also for the material travels. The %age utilization is less compared to %age idle as shown in the Fig 6. Implementation of Milk run for the smooth movement of the material & finished stock was difficult because of the workers of the assembly unit & the workers of the calibration units were located in the same area. The movement of the trolleys & spiders were difficult to move. Both the Work in process (WIP) inventory & finished stock inventory were getting piled up & causing bottle necks. This was the main Lean Manufacturing waste.

b) Final Proposal for the Lean Line Layout

The proposal in Integrated Lean Line Design had the disadvantage of bottlenecks in the process of assembly, increase in idle time, Milk run problems. A Detailed discussion with the Managers & the concerned Engineers are carried out regarding the problems of Milk run, bottleneck issues, space constraints and the considerations for reducing the number of activities & the number of workers and Simulation results are analyzed, compared & a Lean Flow Line Layout is proposed & shown in the Fig 7.

![Fig 7 Lean Line Layout](image)

The layout has the following advantages.

a. The number of workers reduced from 33 to 28
b. The number of stations reduced from 21 to 14
c. Effective implementation of Milk run concept for the smooth flow of in process Material.
d. The distance between the workers are comfortable & calculated based on the prototype fabricated
e. Visual Charts, instruction sheets are displayed to implement Pokayoke fool proofing technique.
f. Kanban tools (withdrawal Kanban & Production Ordering Kanban) are studied. Time Study and Methods Time Measurement (MTM) studies are carried out for optimizing the layout design. The ergonomic principles are suggested for the implementation of the Layout.
g. Milk run provisions are made effectively. The layout & analysis using VISIO 2007 & PROCESS PLANNER Software for the Proposed Layout is shown in Fig 8 & 9 respectively.

![Fig 8 Lean Line Layout using VISIO 2007](image1)

**Fig 8 Lean Line Layout using VISIO 2007**
*(One Piece Lean Line Design with Milk Run)*

![Fig 9 Lean Line Layout using Process Planner Simulator Software](image2)

**Fig 9 Lean Line Layout using Process Planner Simulator Software**
*(One Piece Lean Line Design with Milk Run)*

This Layout has the features of Effective Milk Run activity with a provision for easy Loading & Unloading of the raw Material & Finished product without obstructing the Workers assembly activities. The components in the assembly unit are effectively carried out with the workers located in the internal U area. The finished components are unloaded easily with the provision of the workers located in the external U location & also the calibration activities are carried out with the workers located in the External U area. The bottleneck in the process is reduced.
Further, the Layout has the features of Kanban controlled movement, supermarkets at appropriate places. Withdrawal Kanban & Production Ordering Kanban is prepared & Visual Charts are placed for the easy assembly of the pumps. The Simulation analysis & output result for the Lean Line Layout- One piece flow design for the assembly of the Pump is shown in the Fig 10.

![Fig 10 Simulation Result using Process Planner Simulation software](image)

It is observed that the %age Utilization of the assembly activity is increased & idle time is reduced as shown in the fig. 11. But during the end stages due to the calibration set up there is increase in idle time. This has to be addressed seriously. The layout has the features of “Internal U Lean Line for the assembly Operation” & “External U Lean Line for the Calibration Division”, which will avoid the congestions in the workers movement & reduce the Work in Process (WIP) & Finished Goods inventory.

**Development of Prototype for the Lean Line Design (Wooden Model)**

The proposed Lean line design is fabricated using wooden models. The photographs showed in the Fig 12 depict the various assembly stations. During the Prototype fabrication, Industrial Engineering concepts of Method Study, aspects of easy reach, unloading, loading locations (Palettes), inspection center, racks, Milkrun provisions, easy accessibility are considered.
CONCLUSION

Simulation is carried out using the VISIO 2007 & Pro Model Process Simulator software. The Models are fabricated. The feedbacks regarding the Method Study & Ergonomic aspects are carried out using survey technique. Simulation analysis is carried out & the study suggested the schemes for effective implementation of the Milkrun concept for the flow line design.
REFERENCES

Performance Analysis of AODV and DSR in MANET—A Simulative Study

Gaurav Sethi  
Dept. of Electronics and Communication Engineering  
Lovely Professional University  
Jalandhar, Punjab, India  
e-mail: gaurav257161@rediffmail.com

Manoj Kumar  
Dept. of Electronics and Communication Engineering  
DAV Institute of Engineering and Technology  
Jalandhar, Punjab, India

Abstract—Mobile ad hoc networks (MANETs) are collections of mobile nodes, dynamically forming a temporary network without pre-existing network infrastructure or centralized administration. These nodes can be arbitrarily located and are free to move randomly at any given time, thus allowing network topology and interconnections between nodes to change rapidly and unpredictably. The dynamic nature of mobile ad-hoc networks makes traditional routing protocols unsuitable for MANETs. Routing protocols designed for ad-hoc networks have to fulfill a unique set of requirements. This paper presents the simulative performance analysis of The AODV and DSR protocol used in MANETs. This paper establishes that out of the two protocols used here, AODV is better suited in MANET, as end-to-end delay is increased in case of DSR in normal condition and also with the change in mobility.

Key Words: - MANET, AODV, Otcl, NS-2

I. Introduction

A Wireless ad hoc network consists of wireless mobile nodes. Such a network does not have a fixed infrastructure but nodes perform the networking function by acting not only as a host but also as a router, forwarding packets to other nodes that may not be within direct wireless transmission range of each other. That is why ad hoc networks are also sometimes called multi-hop wireless ad hoc networks [1-3]. Several protocols have been developed under the authority of Mobile Ad hoc Networking(MANET) working group. MANET is charter of Internet Engineering Task Force (IETF). Lots of research has also been done about the performance of ad hoc networks under varying scenarios. Different kind of metrics or characteristics may be used to analyze the performance of an ad hoc network. Different kind of approaches and methodology has also been used. Simulations are commonly utilized especially when analyzing the performance of specific routing protocol [4]. Analytical models have also been developed for use especially in analysis[5] considering a specific performance issue of ad hoc networks in general. This paper is organized as follows: After introduction in section 1, a brief overview of AODV & DSR ad hoc routing protocols is given in section 2&3. Section 4 provides the details of the simulation environment & methodology used in analysis of the two routing protocols. Section 5 elucidates the various performance metrics used for carrying out the comparative analysis of the two routing protocols. Results & discussion are presented in section 6 and finally the important conclusions drawn are summarized in section 7.

II AODV routing algorithm

A. Route Discovery in AODV

Upon sending a data packet, the source node first checks its route table to see if any route to destination is available. If yes, it unicasts the data packet to the next node listed in the route table entry[1]. Otherwise, it starts the route discovery, after augmenting its sequence number and broadcasts the RREQ packet [6]. The broadcast ID, maximum hops count, source & destination node address and source & destination node sequence numbers are put into the RREQ packet. Upon receiving a RREQ packet, each node first checks to see if it has seen the packet with the same source address but same or higher broadcast ID before. If yes, it discards the packet. Otherwise, it first sets up a reverse route entry for the source node, then checks to see if it has a route to the destination in its table (including itself). If a route can be found, it then unicasts a RREP packet containing required hops count back to the source. Otherwise, it re-broadcasts the RREQ after increasing the hops count into the packet. The packet is discarded if its total hops count reaches the maximum hops count specified in it. Upon receiving a RREP packet, the node checks to see if it itself is the source node [4]. If yes, it records the destination, neighboring node address and destination sequence ID and hops count to the destination into its route table entry. If there is no such entry to the destination, or if there is a previous entry with lower destination sequence number, or with same destination sequence number but more hops count. Other wise, it sets up a forward route entry in its route table, then unicasts the packet back according to the previously set up reverse route[1].

B. Route Maintenance in AODV

Upon receiving a data packet, the node checks to see if itself is the destination node. If yes, the packet successfully reaches its destination [1]. Otherwise, it looks up the next node to the destination in the route table and unicasts the data...
packet to next node. If the packet fails to be delivered, the node unicasts the RERR packet back along the reverse route to report unreachable destination [7]. Upon receiving a RERR packet, the node checks to see if itself is the source node. If yes, it removes the broken route entry to the destination in its table that are broken as reported by RERR packet, redo the route discovery or resend the data packet process, if necessary. Otherwise, it unicasts the packet back according to the route specified in the packet, marks the forward route to the destination unreachable.

III. Dynamic Source Routing

A. DSR Routing Algorithm

The Dynamic Source Routing protocol (DSR)[1] is a simple and efficient routing protocol designed specifically for use in multi-hop wireless ad hoc networks of mobile nodes. Using DSR, the network is completely self-organizing and self-configuring requiring no existing network infrastructure or administration. Network nodes co-operate to forward packets for each other to allow communication over multiple "hops" between nodes not directly within wireless transmission range of one another. As nodes in the network move about or join or leave the network and as wireless transmission conditions such as sources of interference change, all routing is automatically determined and maintained by the DSR routing protocol [8]. Since the number or sequence of intermediate hops needed to reach any destination may change at any time, the resulting network topology may be quite rich and rapidly changing. The basic version of DSR uses explicit "source routing", in which each data packet sent carries in its header the complete, ordered list of nodes through which the packet will pass. This use of explicit source routing allows the sender to select and control the routes used for its own packets supports the use of multiple routes to any destination (for example, for load balancing) and allows a simple guarantee that the routes used are loop-free; by including this source route in the header of each data packet, other nodes forwarding or overhearing any of these packets can also easily cache this routing information for future use. The DSR protocol is composed of two main mechanisms [9] that work together to allow the discovery and maintenance of source routes in the ad hoc network:

B. Route Discovery

It is the mechanism by which a node S is able to detect, while using a source route to D, if the network topology has changed such that it can no longer use its route to D because a link along the route no longer works. DSR [1] uses two types of packets for route maintenance: Route Error RERR packets and ACKs. Whenever a node encounters fatal transmission error so that the route becomes invalid, a source receives a RERR message. Source then removes the erroneous hop from all of its route cache entries and selects a new route, or if there are no more available routes, it initiates a new route discovery [4]. ACK packets are used to verify the correct operation of the route links. Route maintenance for this route is used only when S is actually sending packets to D.

In DSR, both route discovery and route maintenance operate entirely "on demand". In particular, unlike other protocols, DSR requires no periodic packets of any kind at any layer within the network. In response to a single route discovery (as well as through routing information from other packets overheard), a node may learn and cache multiple routes to any destination. This support for multiple routes allows the reaction to routing changes to be much more rapid, since a node with multiple routes to a destination can try another cached route if the one it has been using should fail. This caching of multiple routes also avoids the overhead of needing to perform a new Route

The DSR protocol [1] is designed mainly for mobile ad hoc networks of up to about two hundred nodes and is designed to work well with even very high rates of mobility.

IV. SIMULATION ENVIRONMENT

The results reported in this paper are based on the study conducted on the basis of simulation tool NS (version 2), that is an object-oriented, discrete event driven network simulator developed at UC Berkely written in C++ and OTcl. NS is primarily useful for simulating local and Wide area networks, wireless LANS and ad hoc networks. Currently, NS (version 2) written in C++ and OTcl (Tcl script language with Object-oriented extensions developed at MIT) is available. NS uses a Object oriented Tcl interpreter towards the user. This means that user writes the Otcl script, which defines the network (nodes, links), the traffic in the network (sources, destination, types of traffic) and which protocol it will use. NS then uses this script
during simulation. The result of simulation is an output trace file that can be used to do data processing (calculate delay, throughput) and to visualize the simulation with a program called network animator. Nam is a very good visualization tool that shows how packets propagate through the network. Trace graph is a tool used to plot graphical relations between various parameters.

V. PERFORMANCE METRICS

A. End-to-End delay: This implies the delay a packet suffers between leaving the sender application and arriving at the receiver application.

B. Throughput of received Packets: It is the number of received packets received per TIL (Time Interval Length).

C. Packets dropped: It is the number of packets dropped per TIL.

D. Scalability: This parameter evaluates the performance of the network as the network size increases.

E. Mobility: Mobility is the major parameter of an ad-hoc network. Since an ad-hoc network is primarily characterized by its ever-changing topology, so mobility of nodes is an important consideration. Mobility of a node is a function of both speed and movement patterns.

F. Jitter: An unwanted variation of one or more signal characteristics.

G. Sequence Number: Sequence number is a number given to a packet so that data arrive in sequence and not lost.

VI. RESULTS & DISCUSSION

In the simulation set-up used for carrying out the performance analysis of routing protocols, we have considered a total no. of 20 nodes on a duplex link at a bit rate equal to 1Mbps and drop tail of 10ms, packet size of 500 at 0.005 interval using user data gram protocol(UDP).

A. AODV Protocol

a) Throughput of received packets

It is observed that during the complete transmission, reception of packets is consistent throughout as shown in fig.1. a total no. of received packets/ TIL equal to 80 is obtained.

b) End-to-End Delay

This parameter comprises all kind of delay i.e. delay that occurs when the packet is stored in a buffer before the node transmits it to other node, transmission delay etc. Figure 2 shows the results for average end-to-end delay v/s throughput of receiving bits (bit/TIL) for AODV protocol. The route recovery takes a lot of time as shown in fig. 2 below. It is investigated that for 1x10 bit/TIL, an average end-to-end delay of 0.15 sec is recorded.

Figure 2. Throughput Vs delay


c) Packet drop

AODV has good performance because packets are not sent until a route is not found as observed in fig. 3

d) Scalability (No. of node changes)

1. Throughput of received packets

It is observed from fig. 4 that as total number of nodes change, although throughput of received packets decreases but response does not changes considerably. Here, it was observed that a total no. of receive packets/ TIL equal to 60 is obtained on increasing the total no. of nodes.
Ad-hoc networks are characterized by its ever-changing topology due to mobility of nodes. So it has become essential to monitor the performance of protocols when mobility of nodes is varied. When we talk of mobility, it means change in movement pattern and speed of nodes. Thus, we varied the speeds and locations of some nodes during the simulation to increase the mobility of nodes.

1. Throughput of received packets
Fig. 5 shows that the throughput of received packets doesn’t remain constant and even drops to 50% of its value in normal conditions when nodes are considered mobile in the case of AODV protocol. A total no. of received packets/ TIL equal to 40 is recorded.

2. End-to-End Delay
It is investigated that mobility introduces an increase in the average delay as observed from fig. 6. Here an average end-to-end delay of 0.3 sec was recorded for 1x10 bit/TIL.

f) JITTER
It is observed from the figure that jitter of generated packets varies from 20ms to 40ms. The maximum value of jitter goes upto around 50ms at sequence number around 3200.

g) SEQUENCE NUMBER
In a packet, a field of the packet header that is used by the terminating endpoint to determine if the packets arrive in sequence.
Figure 8 shows graph between packet generation time and sequence number of generated packets. It is observed that the generation of sequence number is the linear function of packet generation time.

**B. DSR Protocol**

*a) Throughput of received packets*

It was investigated from fig. 7 that in DSR protocol, the throughput does not remain constant, however a minimum throughput of the order of a total no. of received packets/ TIL equal to 140 is maintained.

**Figure 9. Throughput of received packets**

*b) End-to-End Delay*

Figure 8 shows the results for average end-to-end delay v/s throughput of receiving bits (bit/TIL) for DSR protocol. It is investigated that for 1x10^15 bit/TIL, an average end-to-end delay of 0.8 sec is recorded. Also the route discovery is fast as compared to AODV.

**Figure 10. End-to-End Delay**

**c) Packet drop**

The route discovery is difficult in DSR protocol as compared to AODV protocol. But once a route is established, packet drop rate is less in DSR protocol. Fig 9 shows that number of packets dropped is less as compared to AODV.

**Figure 11. Throughput of dropped packets**

**d) Scalability (No of node changes)**

1. **Throughput of received packets**

It is observed from fig. 10 that as total number of nodes change, initially the throughput of received packets keeps low but it catches up very fast and is able to maintain throughput of a total no. of received packets/ TIL equal to 140-160 on increasing the total no. of nodes. Thus as number of nodes changes, response does not changes considerably.

**Figure 12. Throughput of received packets**
e) MOBILITY SIMULATIONS

1. Throughput of received packets
   It is investigated that throughput for DSR remains same even with the change in mobility as observed from fig.11. The throughput measuring a total no. of received packets/TIL equal to 130 is maintained.

![Figure 13. Throughput of received packets](image)

2. End-to-End Delay
   It is observed from fig. 12 that average delay is higher in case of DSR protocol when mobility is introduced i.e. when nodes move. Here an average end-to-end delay of 0.95 sec was recorded for 1x10^5 bit/TIL.

![Figure 14. Throughput of received bits Vs Average delay](image)

f) JITTER
   It is observed from the figure15 that jitter of generated packets is almost negligible. Hence there is less disturbance in DSR.

![Figure 15 Jitter of generated packets Vs Sequence number](image)

![Figure 16 Sequence number Vs Packet generation time](image)

g) SEQUENCE NUMBER
   It is observed from the graph that the generation of sequence number in case of DSR is in non-linear manner as contrary to AODV.

![Figure 17 Sequence number Vs Packet generation time](image)

VI. CONCLUSION

In AODV protocol, reception of packets is consistent throughout and a throughput measuring a total no. of received packets/TIL equal to 80 is obtained. While in DSR protocol, the throughput does not remain constant, however a minimum throughput of the order of a total no. of received packets/TIL equal to 140 is maintained. The route recovery takes a lot of time in AODV protocol and an average end-to-end delay of 0.15 sec is recorded for the throughput of 1x10^5 bit/TIL. It is investigated that an average end-to-end delay of 0.8 sec is recorded in case of DSR protocol. Also the route discovery is fast as compared to AODV. AODV has good performance because packets are not sent until a route is not found. The route discovery is difficult in DSR protocol as compared to AODV protocol. But once a route is established, packet drop rate is less in DSR protocol. No. of packets dropped is less as compared to AODV. In AODV protocol, as total number of nodes change, throughput of received packets decreases whereas in DSR protocol, initially the throughput of received packets keeps low but it catches up very fast and is able to maintain throughput of a total no. of received packets/TIL equal to 140-160. Thus as number of nodes changes, response does...
not changes considerably. Throughput of received packets doesn’t remain constant and even drops to 50% of its value in normal conditions when nodes are considered mobile in the case of AODV protocol. It is investigated that throughput for DSR protocol remains same even with the change in mobility. The throughput measuring a total no. of received packets/ TIL equal to 130 is maintained. It is investigated that mobility introduces an increase in the average delay. Here an average end-to-end delay of 0.3 sec was recorded in AODV protocol. in case of DSR protocol, average delay is higher and of the order of 0.95 sec for the throughput of 1x105 bit/TIL. Jitter is more in case of AODV leads to the more disturbances during data transmission but in case of DSR jitter is less leads to more smooth performance also sequence number is generated more linearly in case of AODV but once route is established generation of sequence number in DSR is fast as compared to AODV Hence AODV scores over the DSR protocol as end-to-end delay increases in case of DSR in normal condition and also with the change in mobility.

VII. REFERENCES

Development of a Trusted Third Party Time-Date Stamp Server in an IT Security Course
Gordon Romney, National University

Universities have many digital assets such as documents, manuscripts, photographs, images and audio/video recordings for which provenance need to be validated for both scholarly and legal reasons. One highly secure technology that may be used in validating provenance is that of asymmetric cryptography, and, specifically, a digital signature. Establishing when an asset is digitally signed becomes the critical issue as a computer system clock may be easily adjusted. Creating, for university usage, a verifiably accurate system time and date stamp (“TDS”) that can be proved in a legal setting is the objective of this research. Such a service in e-commerce qualifies as one implementation of a Trusted Third Party provider.

A TDS server was designed and developed by students in an Information Technology (“IT”) security course. Located in a secure, centralized facility on campus, it was implemented using open source software operating under Linux. The TDS synchronized with the NIST atomic clock twice a day. In turn, the TDS clock was used by remote nodes to synchronize their system clocks twice a day. A record of all adjustments made to any of the system clocks was archived in a history file that was logically sequenced and digitally signed. A complete history of TDS time from an initial time $t_0$, subsequently, may be established. The same is the case for each independent, autonomous node.

Undergraduate IT students demonstrated their ability to transfer knowledge from theoretical instruction on public key infrastructure to an operating TDS system of significant complexity. The TDS server was successfully interfaced with NIST and a remote node that provides digital signature signing capability for signing digital images of ancient manuscripts. The database that contains the clock histories was implemented in MySQL. This centralized TDS server provided a Trusted Third Party service as it was organizationally independent from any of the nodes or node owners.
Simulative Analysis of operating conditions for Flat amplitude Multiwavelength Brillouin-Raman comb fiber laser

Kamaljit Singh Bhatia
Dept. of Electronics and Communication Engineering
DAV institute of Engineering and Technology,
Jalandhar,Punjab,India

Dr. Manoj Kumar
Dept. of Electronics and Communication Engineering
DAV institute of Engineering and Technology,
Jalandhar,Punjab,India

Abstract: We investigate the operating temperature and pumping power level of Rayleigh-assisted Brillouin-Raman comb laser in a linear cavity in which feedbacks are formed by high-reflectivity mirror. The optimization of Brillouin pump power and wavelength is very crucial in order to obtain a uniform power level between Stokes lines. The Brillouin pump must have a relatively large power and its wavelength must be located closer to the Raman peak gain region. A flat-amplitude bandwidth of 30.7 nm from 1527.32 to 1558.02 nm is measured when Raman pumping level is set to 30dB and operating temperature to 300K. The Q value of 31.3774[ln] dB is obtained for all the Brillouin Stokes lines at data rate of 9.953Gbps.

1. Introduction
Multi wavelength fiber lasers are one of the attractive solutions to support dense wavelength division multiplexed systems. There are two common techniques commonly utilized; external channel filtering and internal channel generation. The former technique is based on slicing a broadband spectrum from a super continuum source [1]. The latter technique is to produce multiple channels using any means of filtering within a laser cavity internally. A physical narrowband filter can be utilized as one of the options in this technique [2]. On the other hand, a group of laser lines can be generated from a seed channel. This approach is commonly known as hybrid-gain configuration which manipulates narrow bandwidth of Brillouin gain in optical fibers. The idea is first demonstrated in Brillouin-erbium fiber laser cavity [3]. Later, it is extended to incorporate the nonlinear Brillouin scattering in Raman fiber laser cavity [4].Since the Raman amplification has wide gain bandwidth, the generation of Brillouin Stokes lines is higher compared to its counterpart of Brillouin-erbium fiber lasers. Brillouin-Raman fiber laser requires a relatively long nonlinear fiber owing to the nature of Raman amplification. In addition, the proposed laser structure in Ref. 4 based on linear cavity in which the signal travels bidirectionally in the amplifying fiber. Therefore, the generation of Brillouin Stokes lines is assisted by Rayleigh scattering. This complicated nonlinear interaction is well explained in the follow-up research work.
Referring to the proposed laser structure, only one high reflectivity mirror is used and at the other cavity end, a virtual mirror is formed taking advantage of Rayleigh double-scattering. Owing to the process of Rayleigh scattering, this virtual mirror has weak reflectivity compared to the physical mirror. Thus, the proposed laser cavity is driven into deep saturation to push Rayleigh component to reach the same saturation level set by the Brillouin components. On the other hand, the distinctive power level discrepancy is clearly observed when the Brillouin-Raman fiber laser cavity is constructed from two virtual mirrors (no physical mirror at both cavity ends) [6]. Later on a linear cavity formed by the high reflectivity element at both ends of the laser cavity was proposed [7]. Gain flattening of single-pump fiber Raman amplifiers is possible by use of non-linear broadening of the pump spectrum [8]. It suggests that pump broadening of a continuous-wave Raman fiber laser tuned at 1455 nm is achieved by propagation in a non-zero dispersion-shifted fiber with small anomalous dispersion at the pump wavelength. The broadened pump spectrum can reach a full width at half maximum of as much as 28 nm. A novel method was proposed for designing multiwavelength pumped fiber Raman amplifiers with optimal gain-flatness and gain-bandwidth performance constitutes a substantial improvement in gain flatness compared to the existing wide-band optical fiber amplifiers [9]. A high Raman gain fiber has been developed that is applicable to Raman amplification over a wide wavelength region [10] We finalise the values of operating temperature and pumping power level for hybrid amplified system. In the research work, the problem of power level discrepancy is successfully resolved. Furthermore, the optical signal-to-noise ratio (OSNR) of the Stokes lines is higher than those achieved from the previous research works. However, all the experimental results obtained with the expense of the amplitude flatness which has been superiorly demonstrated in Ref. 5 and 6. In this paper, we concentrate on the amplitude flatness issue based on the Brillouin-Raman linear-cavity fiber laser. The behavior of Stokes lines that produces flat power level is investigated in details for single-wavelength pumping scheme. In this work, a counter measure to produce flat-amplitude Stokes lines in a wider wavelength range is suggested.
2. Theory

Optical amplifiers are indispensable for realizing a long-distance and large-capacity optical communication system. The optical amplifiers are also effective as a means for compensating for a decrease in light signal intensity in a metro/access system, and are widely used in various optical communication systems.

One of the representative optical amplifiers is a rare-earth-doped optical fiber amplifier such as an erbium-doped fiber amplifier (EDFA) that is mainly applied to a signal band of 1.55 micrometer that is an amplification band of the EDFA. However, in recent years, since a transmission capacity required of an optical communication system has been increasing rapidly, a significant increase in a signal band is required. Therefore, it is difficult to secure a sufficient signal band only with the rare-earth-doped optical fiber amplifier. An optical amplifier applicable to optical signals in a wider band is demanded.

A Raman amplifier makes use of stimulated Raman scattering that is caused by pumping light lead into a silica fiber. Therefore, it is possible to set an amplification wavelength freely by changing a pumping light wavelength and realize a desired gain wavelength characteristic by adjusting an output distribution of pumping light consisting of a plurality of wavelengths. These are characteristics advantageous for an optical amplifier used for broadband wavelength multiplexing transmission. The rare-earth-doped optical fiber amplifier does not have the characteristics.

Various technologies utilizing these characteristics of the Raman amplifier have been proposed. The technologies include a technology for adjusting a gain wavelength characteristic by changing a pumping light output consisting of a plurality of wavelengths arbitrarily (U.S. Pat. No. 6,115,174) and a technology for realizing a flat gain wavelength characteristic by setting a plurality of pumping wavelength intervals properly (U.S. Pat. No. 6,292,288).

3. System Description

The simulation set-up for modeling of Raman amplifier’s operating temperature and pumping level determination method using standard fibers is shown in figure 2.1.
The transmitter section as shown in the figure 2.1, consists of data source, electric driver, laser source and amplitude modulator. The data source is modulated using different data formats at bit rates varying from 2Gb/s to 9.953 Gb/s. The laser is of the type CW Lorentzian with laser center emission frequency 1550nm (193.4145 THz). The amplitude modulator is of type sine square with excess loss of 3 dB. The output of modulator is fed to optical link consisting of two optical spans of varying length having booster, optical splitter and an standard optical fiber. The optical splitter is connected to correlate the signal at input and output of two optical spans using optical spectrum analyzers. The optical amplifiers are of fixed output power type each with noise figure of 5 dB. The output is detected at the receiver by PIN detector with responsivity 0.875 and is passed
through electric filter and output is observed on electroscope. The electric filter is of the type Bessel with -3dB bandwidth equal to 8GHz. The electroscope gives eye diagram, Q value, bit error rate and eye closure penalty. The effects of fiber non-linearities and polarization mode dispersion are also considered in the simulative analysis and Iterations has been taken for power, temperature, pumping and fiber length as shown (Fig:3.2)

<table>
<thead>
<tr>
<th>Power</th>
<th>temp</th>
<th>pumping</th>
<th>LENGTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 1</td>
<td>10</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Run 2</td>
<td>12.5</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Run 3</td>
<td>15</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td>Run 4</td>
<td>17.5</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Run 5</td>
<td>10</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Run 6</td>
<td>10</td>
<td>600</td>
<td>300</td>
</tr>
</tbody>
</table>

**Table1: Iteration for different parameters**

4. Results and discussion

(a)

(b)
Figure 2: At 9.53 Gbps for NRZ raised cosine (a) Power verses temperature (b) Power verses pumping (c) Power value verses length (d) Eye diagram to show different performance indices
### 5. Measurements:

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Parameters</th>
<th>Power Level (db)</th>
<th>Operating Temperature(K)</th>
<th>Optical Length(km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.5 Gbps</td>
<td>20</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>2.</td>
<td>3.0 Gbps</td>
<td>20</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>3.</td>
<td>4.0 Gbps</td>
<td>20</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>4.</td>
<td>5.0 Gbps</td>
<td>10</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>5.</td>
<td>6.0 Gbps</td>
<td>10</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>6.</td>
<td>9.953 Gbps</td>
<td>30</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td>7.</td>
<td>12 Gbps</td>
<td>20</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>8.</td>
<td>15 Gbps</td>
<td>10</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

**Table 2: Operating parameters at different data rates**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.5</td>
<td>26.89</td>
<td>0.011</td>
<td>1e-040</td>
<td>0.03199</td>
</tr>
<tr>
<td>2.</td>
<td>3.0</td>
<td>4.42</td>
<td>0.018</td>
<td>0.00017</td>
<td>0.0063</td>
</tr>
<tr>
<td>3.</td>
<td>4.0</td>
<td>3.59</td>
<td>0.014</td>
<td>0.0037</td>
<td>0.0019</td>
</tr>
<tr>
<td>4.</td>
<td>5.0</td>
<td>20.46</td>
<td>0.032</td>
<td>1e-040</td>
<td>0.0096</td>
</tr>
<tr>
<td>5.</td>
<td>6.0</td>
<td>2.0</td>
<td>0.0045</td>
<td>0.0022</td>
<td>2.267e-005</td>
</tr>
<tr>
<td>6.</td>
<td>9.953</td>
<td>31.2376</td>
<td>0.0045</td>
<td>1e-040</td>
<td>0.01772</td>
</tr>
<tr>
<td>7.</td>
<td>12</td>
<td>2.42881</td>
<td>0.0228</td>
<td>0.0081</td>
<td>8.188e-005</td>
</tr>
<tr>
<td>8.</td>
<td>15</td>
<td>2.0</td>
<td>0.027</td>
<td>0.022</td>
<td>2.26e-005</td>
</tr>
</tbody>
</table>

**Table 3: Simulation results at different data rates**
Conclusion
In conclusion, we have investigated the amplitude flatness of Rayleigh-assisted Brillouin-Raman linear-cavity fiber laser. High-reflectivity mirrors used at both ends of the laser cavity create strong oscillation of cavity modes around the Raman peak gain. The optimization of power and operating temperature has great influence in determining the uniformity of the Stokes lines amplitude. The flat amplitude bandwidth is also governed by the number of Raman pump wavelength used in the laser structure. The maximum Q value of 31.3774[lin] is obtained when the laser cavity is pumped by single wavelength pumping scheme at 1435 and 1450 nm with total pump power of 30 dB. Owing to the high-reflectivity feedbacks, the average OSNR value of 17 dB across the flat-amplitude bandwidth is successfully measured. The flexibility and scalability of Raman pumped laser cavity has great promise to provide multiple channels in larger bandwidths.

References

**GridARTCA**: A collaborative technology environment utilizing Cyberinfrastructure - A preliminary report

Kevin Franklin¹, Santiago Nunez-Corrales², Jose Castro², Peter Bajcsy¹, Alvaro de la Ossa², Thomas MacCalla³, and Arun Datta³

¹National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, ²National Collaboratory for Advanced Computing, Center for Advanced Technology Studies, ³National University Community Research Institute, National University, San Diego

ABSTRACT: Application of advanced computer technology can be a boon to the economy of any country. The Costa Rican Government has emphasized the importance of such technologies and a newly emerged committee, which is the Advanced Research and Technology Collaboratory for the Americas (ARTCA) has committed to support and oversee the effective collaboration between the industry and the research institutions in the interest of the economic growth in Costa Rica. This is a cooperation program for applied research and development of advanced technologies, in particular supercomputing technologies for the sciences, engineering, humanities, arts and social sciences, which aims at expanding its scope throughout the Americas. The objectives of ARTCA are interdisciplinary, interuniversity and international cooperation through the continuing development of the human resources focusing on researchers and students from universities and R&D centers.

This cooperation among multi-institutions is envisioned to be achieved through cyberinfrastructure that would integrate resources from collaborative partners worldwide based on the utilization of global science and technology blending innovation, which would give rise to a successful regional strategy based on the application of computing to problems led by the National Center for Advanced Technology Studies (CeNAT) in Costa Rica. Among the projects that have already been defined, some of which are already under execution, are the development of visualization and image analysis applications for the aerial land images of Costa Rica obtained from the CARTA 2003 and 2005 missions, coordinated by CeNAT and with support from NASA. Another priority is to utilize the advanced technology for solving the health related problems involving dengue causing parasites.

The gridARTCA system will support the cyberinfrastructure by integrating all the technology resources through the utilization of Giga bit network for handling massive amount of data. This system would be developed using open source software that will enable the users to collect data through multiple computerized means, store data in a relational database, and report related statistical and tracking information on dengue affected patients, as for example in a case of dengue epidemiological study. This system will also provide web based data entry forms for various users –in the example case for healthcare service providers- and help significantly improve data workflows. The software system of gridARTCA also will determine and allow a user to view his/her certain domain-specific related reports and analysis, following multiple standards and ensuring compliance with them. This system would also support (a) the local development of science and technology, (b) existing collaborations with institutions in the US and Europe in both areas, (c) research on the existing local data (d) scientific resources and (e) compliance with a national strategy for the advancement of Science and Technology, namely, the XXI Century Strategy.
Serialized Tokens as a Means of Information Access Control

Miles Romney IV (CTO, Matchbin, inc., miles.romney@matchbin.com)
Todd H. Romney (CEO, iMrkr, inc., todd.romney@imrkr.com)
Gordon W. Romney, Ph.D. (Department of Computer Science and Information Systems, National University, gromney@nu.edu)

Abstract

With the explosion of information portals on the web, and the resultant transition of ordinary users into information providers as well as information consumers, comes a host of challenges surrounding the control of access and the selective distribution of information. How can the right information be delivered to the right audience? How can sensitive information be protected from all but a select target audience? How can different forms of the same information be efficiently prepared and stored for delivery to different audiences? How can access-controlled information be taken beyond the web, into the physical world?

One solution to this problem, which provides efficient answers to each of these questions, is the use of serialized tokens to control information access. These tokens can take either digital or physical form, and provide the bearer with an access identity that can be specifically configured to unlock any information group, or even a select subset of one or more information groups. These tokens go beyond assigning users a single monolithic access definition, and allow for a many-to-many user-publisher-to-user-consumer relationship where each user has complete control over the information – and its form – that she shares with other users. In digital form, a user could trigger the automated delivery of a digital token via e-mail or any other communications medium. In the physical world, a person can simply hand-deliver a serialized physical token, which he can configure at any time with focused access to specific information.

1.1 Compulsive Confessions, 21st Century Style

Human nature hasn’t changed in a thousand years: so goes a popular argument. We’re still violent, we still cry for a sad story, and we still pass along spicy rumors about the neighbors to the north to the ones on the south. The difference is, we’re now in the 21st Century, which means we have the tools to throw these habits in a far wider net. Improved weaponry makes anger more dangerous; a million people gather in movie theaters across the world to see the latest chick flick; and all the latest gossip is spread across the pages of blogger.com.

The only thing that has changed, it seems, is technology, and the technology that has allowed us to tell tales to millions instead of half-a-dozen is the World Wide Web. We’re all compulsive confessors, says popular pundit Tucker Carlson [1]. The World Wide Web brings countless efficiencies, among them, unfortunately, the very efficient power to cause ourselves material harm by publishing our confessions to the world.

The dangers of global confession, happily, have been well identified and well publicized. Various efforts have emerged to educate the public on the potential dangers of publishing personal information online [8]. Many college campuses have incorporated warnings on the potential ills of web publishing into their orientation materials, in an effort to demonstrate to new students the starkly public nature of online publishing. “The particular focus is the public nature of this,” said Tracy Tyree, dean of student life at Susquehanna University. “That seems to be what surprises students most. They think of it as part of their own little world, not a bigger electronic world.” [2] And a “bigger electronic world” it most definitely is—MySpace alone boasts over 90 million registered users, and is the second most visited site on the Web [3]. And MySpace consists entirely of user generated content, often of a personal nature, and without any privacy tools for controlling access to a user’s published information.

The dangers have been identified, but no robust solutions have been fielded. Some services offer binary, or at best, trinary, security solutions that allow for defining access to a complete body of personal information as “visible”,

302
“invisible” or “minimally visible”, but none of these solutions offers anything close to granular control over individual datum in a many-to-many online environment.

1.2 The State of Information Security on the Web

MySpace is the highest trafficked social networking application [3], but it is hardly the only security offender in the social networking space. Facebook attracts more than 20 million unique visitors per month [4], and offers an array of tools comparable (or even superior) to those at MySpace. Facebook has tried to protect the information of its users by using a trinary visibility system for full profiles (a full profile can be visible, invisible, or minimally visible), but it still publishes user data in public directories, and provides no ability for a user to restrict select information to specific viewers. A friend is a friend is a friend, to Facebook—what’s the difference? Whether at MySpace, Facebook, LinkedIn, Ning, Friendster or Orkut, the state of privacy is largely the same, if it is offered at all: an “on”/”off” switch with no nuance of control.

The pitfalls of unprotected information have become frighteningly clear. The harshest lessons in the dangers of open-information have been dealt to the underaged. Speaking of MySpace, comedian Demetri Martin said “On the downside it's loaded with sexual predators; On the upside, it's loaded with sexual prey.” Disturbing, but certainly telling. The popular culture has evidently accepted the idea that MySpace is a dangerous place to be—but no-one has yet provided a robust solution to the problem. As recently as January 23 of 2008, a pair of young New York girls were fooled by an adult couple posing as teenagers, and lured into sexual abuse [5]. MySpace has responded with plans to restrict usage to users who declare themselves to be under 18, and to set-up a program of keeping parents informed. Half-measures, it would seem, for a problem of this magnitude.

Beyond the issue of sexual predators, though, lie a host of other difficulties. There is the ever-looming threat of identity theft, or the risk of an employer finding and reading material damaging to an employee’s professional standing.

Sharing information has clearly become a national pastime. But who we share information with is every bit as important as what information we share. No matter the information, and no matter the context, it fundamentally comes down to this: the guy I want to give my business card to may not be the same guy I want to share my pumpkin bread recipe with. Where information privacy on the web is largely two-dimensional (what and when), an effective model for information distribution with protection will be four-dimensional: what, when, to whom and from where. Specifically, what information are we sharing, when are we sharing it, with whom are we sharing it, and where does it come from (is it associated with and parented by an information subset)?

Beyond social networking, hundreds of other applications exist that involve communicating information from one person to another, or one person to many specific people, or even many specific people to many other specific people. Much of the time, this involves data duplication—send an e-mail to fifty people, and fifty copies are made and delivered individually. An ideal model would eliminate the need for data duplication, but support the creation of referential subsets that react differently to the what, when, to whom and from where questions.

Whether the information being communicated involves “my favorite foods” in a social context, or product marketing data in a sales environment, or even student grades in academics, it can all be distributed securely and specifically through a system of serialized tokens, each assigned to a four-dimensional permissions definition.

2.1 Four-dimensional data access

Security and data integrity both figure centrally in any data storage and transmission (DST) solution. A scalable security policy can be wrapped around any datum, which answers the questions of what, when, to whom, and from where. A security policy, in this context, is an actual object structure that wraps a datum, defining the conditions under which it may be accessed. Facebook and others offer what they call a “security policy”, but it is simply a statement of best practices, and does not address actual methods of securing data. It is interesting to note that the Friendster security policy references the recommendations of the EU Directive, discouraging the sharing of potentially compromising information. The EU Directive is more security conscious than any analogous offering in the United States.
2.1.1 What

Clearly, the datum itself is the focal point of any security policy. The answer to the what question, however, cannot be taken for granted if data integrity is to be preserved. In the context of database design, data integrity can be considered at risk if a single datum exists in multiple locations. Unless foreign key relationships are impeccably implemented and maintained, data duplication means that one version of a datum may change while the other remains the same, throwing the accuracy of either version into question. [6] The goal in any DST model should be to maintain a single data source, and to reference that source as needed in other applications and contexts. Therefore, every datum must be properly keyed and indexed for use in associative chains.

2.1.2 When

The second in a basic two-dimensional model of DST, the when question is primitively answered with the moment a datum is inserted into the DST pipeline. When a user clicks the “publish” button, the when question is addressed.

More precise control of publication timing has been recognized as desirable, and publication timestamps are now available in most content management systems. It is further helpful to include not only an availability date, but also an expiration date as part of a security policy.

2.1.3 To Whom

The to whom factor is the real entry point in robustly securing a DST pipeline—it is also the first element almost exclusively unavailable in most social networking systems, though it has certainly been implemented (though in a limited fashion) in many other online DST systems.

A primitive implementation of the to whom question might be, for example, choosing to distribute through an exclusive mail list that only delivers to a select group of individuals, rather than posting information to a blog. Another example might involve tethering information to a specific proprietary identity, such as sharing a Google document with a specific e-mail address, or sharing an encrypted note on www.stiki.es with another www.stiki.es username.

The weakness in all of these examples is that access is being restricted based exclusively on proprietary online identity, assuming that the desired recipient, respectively, is a member of your mail list, has a Google account, or is registered with www.stiki.es. Three potentially problematic scenarios immediately present themselves.

• What to do if information needs to be shared agnostically, with no knowledge or regard for which online communities the recipient has embraced?
• What to do if information needs to be shared anonymously, with neither the giver nor the receiver being given personal information about the other in process of transmitting the required information?
• What to do if information needs to be shared or referenced in an offline context, with no respect for online identities?

Each of these situations frustrates current DST security systems that rely exclusively on associating a security policy with an online identity. The first can largely be addressed through any of a variety of universal ID systems, such as OpenID or even Microsoft Passport, but as open as they purport to be, they still represent a fundamentally closed system requiring recognized credentials.

Fundamentally, the rules that govern the sharing of information should not be so different from those that guide software development. Data should be treated as discreet objects, capable of interacting with other objects in predefined ways. Indeed, the “M” (model) in the “MVC” (model-view-controller) philosophy is based on this very idea. A security policy should wrap a datum object, but should not pigeon-hole it. An effective four-dimensional security policy will act as an inter-object interface, describing the data’s availability, and will not prescribe (and therefore limit) its environment. The to whom question, therefore, must be answered in a discreet way: “the datum X object is available to the datum Y object”; and not as a dead end: “the datum X object is available to Bob Jones at Gmail.” Anything less is far from extensible, and cannot universally service information sharing needs.
2.1.4 From Where

The *from where* question is directly associated with the *what* question, and the *what* question will often refer to *from where* in determining its answer. The principles behind *from where* are data integrity and storage efficiency. Integrity is at risk of breaking down when duplicates of the same datum are stored, and clearly, it is inefficient to store duplicates (especially when the datum is not a binary value, but say, a 200 GB video file).

The initial temptation when defining customized security policies which define the *what* and *when* questions identically, but which answer the *to whom* question differently, is to create individual instances of the datum, each to be wrapped in its own security policy. E-mail distribution is a good example. Every time Aunt May finds a new irresistible cat video, she attaches it to an e-mail and sends it to her 80 closest friends, which then creates 80 copies of a 2 MB video, and transmits them to all recipients regardless of which ones have any interest in actually viewing the attachment. Of course, many online services *have* done this right. YouTube is a prime example of effective referencing. You may find a YouTube video on 50 different websites, but each instance is referencing the same video object.

This discipline begins to break down when publication needs start getting granular. On MySpace, for example, what happens if I wish to present a different subset of information to different groups of people? I may want a “Miles Romney” profile to share with friends, a “MechMuse” profile to describe my SciFi publication, or a “Moonroot Studios” profile to represent my independent record label. All of these likely have a great deal of common data. I may want to share location information commonly between them, or post video and audio, some common and some not. Currently, whether I’m on MySpace, Facebook, Ning or Orkut, this scenario would involve my duplicating data.

The *from where* question now gets interesting. In a perfect world, I would be able to define all of these elements—from first name to last name to favorite food to birthday photos—as data objects, capable of being referenced and associated in parent-child relationships. In the earlier example, my “MechMuse” and “Moonroot” profiles would all borrow some subset of information from the “Miles Romney” profile. If I then updated my address on the “Miles Romney” profile, the changes would cascade through my other profiles, because they all simply represent collections of information subsets, comprised of discreet data objects each with an agnostic security policy.

3.0 Referencing Security Policies with Serialized Tokens

The objective, then, is to create a security policy capable of behaving agnostically, but which effectively interfaces with other security policies, triggering the availability of data.

Minimizing data redundancy and preserving the power of granularity in defining access control necessitates describing a security policy for each datum. But data is rarely transmitted in isolation, but rather, in functional groupings. For example, it is rare that you would give someone your middle name, without your first name, last name, and maybe your phone number. Hence, data with common security policies (common in *when*, and *to whom*, though not necessary *what* and *from where*) is grouped into functional subsets, and we’ll simply assign each of those subsets a serialized (but non-sequential) number; we’ll call this number a *token*.

Now, any time a data subset encounters its valid token, the security policies of its constituent data objects will be commonly satisfied, and the information will be made available. The advantage to this model is that it solves each of the three problems referenced earlier, in relation to online-identity-based solutions.

- Information can be shared between any platform, without regard for current registration or membership status of the recipient
- Information can be shared completely anonymously, with neither the giver or recipient being provided any personal information (apart from that actually included in the data subset)
- Information can be shared and referenced offline but still securely by simply hand-delivering a copy of a data subset token
The real-world applications of this arrangement are many and varied, especially when tokens are made available pre-printed, as plastic chips or even business cards. In the social networking space, a guy could pass along a physical token to a girl he meets clubbing, then go home and associate that token with a profile meant specifically for her, with photos of his dog and a blog entry for her eyes only. In business, a salesman could pass along a physical token to a promising sales lead, which he subsequently associates with those product brochures and testimonials most relevant to the sale. In a university setting a physical token could serve as a low-cost student ID, granting a student access to specific information or services wrapped in a four-dimensional security policy. Possession of the token could be combined with a password for additional security, and could even be associated with a student photograph stored server-side. Alternatively, the token could promote a specific university group or school. Online, a simple e-mail invitation could be delivered with an embedded URL containing an applicable access token; the user experience would be identical to many current invitation systems, but would support cross-platform communications, not requiring registration by the recipient, or even permanent use of a specific e-mail address as authentication.

3.1 Database Implementation

A variety of accommodating database architectures is possible, each, no-doubt, with its advantage. The most extensible would likely avoid defining any data types or labels (for the what) in the schema itself, in order to offer as much flexibility as possible to the interface in creating new data kinds and categories. Such a solution may not be optimally efficient, however, in an environment where the nature of the stored data is largely known from the outset.

The authors have implemented a four-dimensional security policy DST system, including token referencing, in a social networking system at www.fobfire.com. MySQL was used in this example, though the concepts described can be applied to any database or other data storage engine. While allowances have been made for free-form user-generated content of an unprescribed nature and context, the system is primarily built around a collection of data called—predictably enough—a profile. The profile represents a defined information subset, where all of the included data has the requisite commonalities in security policies (when and to whom).

In this example, a system has been devised which both preserves efficiency in the database, and allows the data to be accessed by the application as discreet objects, in keeping with the popularized principles of the data abstraction layer [7]. A table has been created to store profiles, and their constituent data. Figure 3.1a illustrates the table structure. Figure 3.1b represents a more theoretically elegant solution, that is however less efficient in an environment where the same information subset will be required almost always, and where database calls need to be minimized in order to optimize performance.

In the case of figure 3.1a, a profile is made up primarily of predefined data fields, each value with a state definition which answers the from where question. Every profile may be associated with a parent profile from which it will inherit values. These associations are theoretically limitless—the "first name" value could be defined in one profile A, with profile B inheriting it. Profile C could then be created which inherits from profile B, which is itself inheriting from profile A. When profile A is updated, both profiles B and C would reflect the change. If profile B were modified to no longer inherit from profile A, but to store its own value, then profile C would inherit that change from profile B, and neither profile B nor C would reflect the value of profile A.

The what and from where questions have therefore been answered. But the when and to whom elements have yet to be addressed. For this purpose, in the fobfire.com example, a separate table has been prepared, called tokens, and is described in figure 3.1c. The purpose of each data field is as follows:

- **profile_id**: Associates a token with a specific information subset, in this case a profile
- **number**: A human-readable serialized (but non-sequential) number identifying the token; this value has been defined as a variable character because it is base-32 encoded
- **availability_timestamp** and **expiration_timestamp**: Datetime value defining the validity period of the token

A combination of the profiles and tokens tables defines all four facets of the four-dimensional security policy.
3.2 Data Models in the Application

The database implementation is concerned, as it should be, with efficiency in both storage and performance. Thanks to the concept of data models in the application, supported by most application frameworks, the database need not be organized according to a strict hierarchical or intuitive system.

The application itself must be able to interact with data as discreet objects, and cannot be concerned with concepts like data inheritance. If the token gatekeeper is satisfied, the application simply calls on any needed datum, ignorant of where it originates, and how it is constructed.

In the fobfire.com example, a strict MVC (model-view-controller) architecture was maintained through the use of the Ruby on Rails (RoR) application framework. Using the data model capabilities of RoR, aggregating values from multiple database tables and arranging data inheritance becomes a simple matter. RoR’s database abstraction engine, ActiveRecord, automatically creates a data object following every database query, with access methods corresponding to each data field. Additional methods may easily be created on the fly, and owing to the extensibility of the Ruby programming language, existing methods may cleanly be extended.

To effectively use the information stored in the database to create an inheritance-aware object, we need only add a single additional method, whose execution will follow every read database query. It might look something like figure 3.2a, which simply extends the default method for each data value, passing each request through an “inherit_value” process which evaluates the datum’s “state”, and returns a native value, and inherited value, or an empty value (in the case of an “exclude” state).

3.3 User Experience

Now that the database has been interfaced with the application, the application needs an interface to the user. The fobfire.com example was intended to leverage a robust and granular security system to allow users in the social networking space to create many profiles, targeted at many different people, and to share information between them.

In the fobfire.com example, a user creates a single “master” profile which contains all the information he may conceivably wish to share with other people. That master profile itself is never shared, but is used as a starting point for all other profiles. Multiple child profiles can then be created, which share as much information from the master profile as desired. A child profile may also be configured with its own custom values, and can include a variety of additional information such as links, favorites, video, audio and other attachments. Figure 3.3a illustrates how a user might configure data inheritance.

Figure 3.3b represents one possible solution for associating tokens (which the fobfire.com implementation calls “fobtags”) with profiles. A user simply hands out a fobtag (token) either physically or via e-mail, then visits the management tool, clicks on the picture of the fobtag he gave away, and assigns it a profile. The profile assignment is only a starting point, however. The user may specifically alter any of the information being associated with any fobtag, to meet the needs of any social or business situation.

4.0 Platform Agnosticism

The fobfire.com example is an effective demonstration of how a four-dimensional security model can be applied to a data storage and distribution system, and integrated with a real-world application such as social networking. One of the major advantages to a token-based system is its ability to maintain platform agnosticism. The next step in applying this system may be the creation of an open specification for four-dimensional security policies, allowing data from disparate sources hosted in a variety of environments to commonly interact based on association with tokens. The fobfire.com example offers an initial foray into this effort by allowing tokens to be associated with assets from other applications. However, in this regard the tokens behave more as aggregators of publicly available information, and less as secure gatekeepers for sensitive information. The application of a common security specification in a variety of applications would allow for a leap in the security (and availability through secure aggregation) of many data from many environments.
As the dangers of compulsive confession become more and more widely realized, and material damages mount, the clamor for a means of granular information control will only escalate. A token-based, four-dimensional security model is an effective and efficient solution, and is capable of being applied to almost any application model.

Figures

**Profiles**

<table>
<thead>
<tr>
<th>id</th>
<th>integer</th>
</tr>
</thead>
<tbody>
<tr>
<td>parent_id</td>
<td>integer</td>
</tr>
<tr>
<td>first_name</td>
<td>varchar</td>
</tr>
<tr>
<td>first_name_state</td>
<td>enum(custom, inherit, exclude)</td>
</tr>
<tr>
<td>last_name</td>
<td>varchar</td>
</tr>
<tr>
<td>last_name_state</td>
<td>enum(custom, inherit, exclude)</td>
</tr>
</tbody>
</table>

Figure 3.1a

<table>
<thead>
<tr>
<th>Profiles</th>
<th>Profile Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>integer</td>
</tr>
<tr>
<td>profile_id</td>
<td>integer</td>
</tr>
<tr>
<td>datum_label</td>
<td>varchar</td>
</tr>
<tr>
<td>datum_value</td>
<td>varchar</td>
</tr>
<tr>
<td>datum_state</td>
<td>enum(custom, inherit, exclude)</td>
</tr>
<tr>
<td>datum_parent</td>
<td>integer</td>
</tr>
</tbody>
</table>

Figure 3.1b

**Tokens**

<table>
<thead>
<tr>
<th>number</th>
<th>varchar</th>
</tr>
</thead>
<tbody>
<tr>
<td>profile_id</td>
<td>integer</td>
</tr>
<tr>
<td>availability_timestamp</td>
<td>datetime</td>
</tr>
<tr>
<td>expiration_timestamp</td>
<td>datetime</td>
</tr>
</tbody>
</table>

Figure 3.1c
References


There Exists a Critical Need for Role-based IDs in a University IT Infrastructure in Order to Meet Security Best Practices

Gordon Romney, National University

This research illustrates the importance of implementing Role-based IDs in compliance with properly established Security Best Practices in order to protect faculty and students as users of a university Information Technology (IT) infrastructure.

Establishing and enforcing Security Best Practices in a university IT environment is a primary responsibility of the university as an employer of student employees. After all our rigor in creating noteworthy pedagogy and dedicating years in the development of an industry-qualified IT student it is a tragedy to place our student IT employee at risk by following sub-standard Security Best Practices. Frequently, this occurs because the University IT infrastructure pursues what is “industry-declared” as an “ideal” objective – namely, Single-Sign-On (SSOn). The frequent approach is to insist on a single ID-Password authentication be given to everyone, and erroneously declared to be SSOn. This fallacious authentication methodology is common in both a university as well as an industry setting.

The proper Security Best Practice is to require a Role-based ID and correlated password. Consider the case of a student employee who has an IT administrative responsibility to change passwords. Good Security Best Practice would require the student to have a student-role (type “S”) ID and password, and as an IT administrator a (type A”) ID and password. When in class, in lab, on campus and around WI-FI hot spots the student would only use her type S authentication. This preserves the type A authentication for use in secure, non-WI-FI, environments when performing administrative functions. The type A authentication is never placed at risk, nor is the integrity of the student-administrator placed at risk.

A case study of an IT student, who was restricted to a single ID-password to be used in both S and A functions, is reviewed in the context of a compromised ID and password that caused the student to be fired from her IT employment, and threatened the student’s graduation and future employment opportunity in the IT industry is reviewed. The steps taken by security professionals to show how a Man-in-the-Middle attack can occur in a WI-FI environment helped avoid damaging the career of this specific student, unprotected by her employer.
The Globally Distributed Ethernet Switch
Barry Brueske, iNetwork Inc., bbrueske@iNetwork-west.com

Abstract

With your IT staff laying out plans for implementing Virtual Servers, massive storage arrays, and disaster recovery sites, wouldn’t it be helpful to be able to link these sites with a truly reliable, active-active multiple redundant, high-speed Layer 2 Ethernet network? And to be able to have a truly “distributed switching” system that can act like, and be managed like, a single, large Ethernet switch? A system where Spanning Tree Protocol doesn’t shut down about half of the bandwidth you bought? Where inefficient VLAN routing (because one location “owns” the VLAN) doesn’t waste a big chunk of the bandwidth? That’s not science fiction – that’s Raptor Adaptive Switching Technology - RAST – and it’s available now.

Keywords: Ethernet, Switching, Distributed, Link Aggregation, Resiliency, Fiber, Spanning Tree, VMWare, SAN, RAST

Please refer to the diagram below. These six locations could be six buildings in a campus, connected by multimode fiber links – or single mode fiber links where they exceed 300 meters (or six continents – more on that later).

All links operate at 10 Gigabits full duplex and all are active paths in a “virtual backplane” within the RAST cluster, in this architecture providing a 20 Gigabit full duplex dual ring with multiple crossover points. Since the 12 individual ER-1010e switch elements bind together into a single logical switch, they all share a common configuration and a common management interface. While Layer 3 switching is supported at the edge of this cluster, it can operate entirely at Layer 2, facilitating virtual server interconnections between multiple locations. Indeed,
VMWare VMotion and distributed ATA on Ethernet redundant storage solutions require a Layer 2 network to span locations. No ordinary “stacking” switch product can support multiple locations like this; no multiple standard Ethernet switch solution could provide the single-switch functionality, or operate in a topology like this without shutting down about half the bandwidth to avoid Ethernet loops.

This is a system that provides the Layer 2 WAN environment that many have been seeking for distributed virtualization. Please see the diagram below:

With RAST, you can have a resilient, high-speed, Layer 2 remote connection without having to deal with Spanning Tree Protocol or other loop avoidance systems that waste available bandwidth and cause other issues. These VMWare resources, as well as the new ATA on Ethernet storage devices, simply cannot talk over Layer 3; they must have a Layer 2 environment. Until now, this created issues in physically separating primary and backup sites, but RAST provides a solution that is attracting users.

The Raptor Ethernet switching system is highly decentralized; switching decisions in this system, whether Layer 2 or Layer 3, are made by intelligent chips at the edge. These highly integrated circuits are aware of all of the ports, VLANs and subnets within the cluster. If a destination is within the same unit, the packet is simply switched to that port within the unit. If the destination is on another unit, the packet is sent to the RAST logic. Packets that must be sent from unit to unit are encapsulated into the RAST protocol, which, rather than shun looped topologies, uses loops for active-active redundant paths, allowing the frames to be directed to the specific destination unit and port by the most efficient available path. One can best picture the RAST system as a multi-unit modular Ethernet switch with a backplane that has been stretched out to reach different cabinets, different rooms, and different cities – with different interface cards and power supplies in each location. In fact, the RAST protocol is analogous to the superframe protocols used on some chassis-based switch backplanes.

The RAST connections operate as a geographically distributed redundant-path backplane. There are multiple paths, and they contain loops, but packets do not loop around within this architecture because the protocol here is not Ethernet, but RAST – so there is no need for anything like Spanning Tree Protocol here. Instead, RAST uses a Shortest Path First algorithm to determine the best path to the desired destination, and sends the packets out the appropriate RAST port towards the destination port. A unit usually will have three RAST ports to choose from, as there are three ports on the ER-1010e RAST module – but two RAST modules could be installed into an ER-1010e (for six RAST ports) to support more complex architectures.
This is a system with no single point of failure. Any devices at any location can be dual-homed to different individual units using the appropriate technology for the device, either Layer 3 or Layer 2. Because RAST is a backplane technology, both Layer 2 and Layer 3 connections and applications can be transparently transported simultaneously. IP-based dual homing is supported for routed dual-homing, and for Layer 2, cross-unit port-channel or Link Aggregation Control Protocol (IEEE802.3ad) is supported. Should one ER-1010e unit go down, the connectivity is maintained, even if one ER-1010e unit in each location goes down. Should a RAST link fail, frames will be sent via another path. In fact, the RAST system will withstand multiple link failures with minimal disturbance; the worst case would be three link failures that isolate one of the ER-1010e units – but then, since the critical elements should be dual-homed, their connectivity to the remainder of the network is still assured.

Management of the RAST system can be in-band, out-of-band, or both, depending on configuration choices. By default, in-band management can exist on any of the ER-1010e’s in the RAST network and can be accessed from any port in the same VLAN as the management interface. But configuration choices can limit management to particular units, say the two units in a given location. If out-of-band management is chosen, that can be done via CLI on the serial port or via GUI on the special front-panel Ethernet port on the manager unit – and in this case, management should be confined to a particular location. A front-panel alphanumeric display will announce which unit is the active manager. Management could also be done from an SNMP management platform; MIBs are provided and an open SNMP management platform is available from Raptor – Raptorview.

As mentioned previously, managing the entire RAST system is like managing a single switch; VLANs and routing interfaces are created once and exist throughout the system. That bears more explanation: VLANs are created as in any layer 2 switch, but are “natively” shared over the entire Distributed Switch Fabric, simplifying configuration and greatly extending the reach of the VLANs without using routing protocols. This allows VLANs to traverse long links and be native in remote sites, similar to VPLS, but with less latency, lower cost and an easier setup. Since no particular unit “owns” any VLAN, VLAN-associated packets are not sent to a “VLAN owner” for switching or routing as on other systems, but routed by the ingress switch chip to the destination switch chip via the shortest path, whether on the same unit or via RAST. IP-based Layer 3 switching is likewise performed by these switch chips, and packets are sent via shortest path to their destination, without needing to be sent to the manager or to a routing engine as on some other systems.

In this fiber RAST scenario, the RAST interface will run at 10 Gigabits to allow using standard 10 Gigabit Ethernet fiber interface components; standard, small form-factor XFP transceivers are
used. Because they are active-active paths, the dual location to location links actually do provide a dual redundant ring with true 20 Gigabit full-duplex throughput per link. A copper RAST interface is also available for local use.

And this brings us to another, important, feature. Because standard 10 Gigabit fiber link hardware is utilized, the distances could be much greater than mentioned above. While we discussed the above network spanning six buildings in a campus, they could also be in six different cities in a metro area. Indeed, current XFP technology supports distances up to 120 kilometers (75 miles), so these locations could each be up to 75 miles apart on leased or private dark fiber. As new transceiver technology becomes available, this distance may become even greater – there is talk of the next step in XFP technology being 200 km – 125 miles. But that is not the distance limit for RAST.

By utilizing DWDM lambdas, these locations could be virtually anywhere; instead of six buildings on a campus, they could be six buildings in, say, San Francisco, Chicago, New York, London, Paris and Tokyo. That would be a truly globally distributed single switch!
Abstract: The study discusses the various aspects of technological innovations that can be incorporated in a tyre manufacturing industry in order to improve the quality of manufacturing. If we look at the value stream of a process, we see that it is customers who actually drive our business. In the manufacturing sense, we therefore need to provide the necessary production responses to satisfy and exceed those expectations by adding value, quality and performance in all that we do. The most effective way of adding value is to have a continuous determination to eliminate waste across the supply chain and thus maximize the value stream: easy to state, difficult to deliver. So, where do the principles, processes and reality of Total Productive Management (TPM) come into play to achieve the goal of a "Totally Productive Operation"? In this context, a study of Why-Why analysis, Cause & Effect analysis, Kaizen techniques were carried out to identify the causes for the various breakdowns in 4 roll calendar Mill, 4-Roll calendar Warm-up mill, Warm-Up Mill Over Head Conveyor. However, in this paper, the study & method plan to be adapted generally to any mill is discussed. Further, elimination of wastes and breakdowns are reduced by the introduction of the following modern tools of management namely Fuguai, Kaizen, One point lesson technique, GTF, CLIT.

Keywords: Fuguai, Kaizen, CLIT, One Point Lesson (OPL), Good to Find (GTF),

1 INTRODUCTION

Lean Engineering & management is gaining momentum in the processing industries in order to improve the process of manufacturing. In this connection, innovative ideas of reducing the wastes & time in manufacturing are being implemented in the organizations. This study is a practical study carried out in a tyre manufacturing industry in Mysore where one of the main issue of Total Productive Management (TPM) is taken up as it is a critical adjunct to lean manufacturing. TPM is also one of the world-class proactive maintenance/lean manufacturing strategies. If machine uptime is not predictable and if process capability is not sustained, we cannot produce at the velocity of sales. One way to think of TPM is "deterioration prevention" and "maintenance reduction", not fixing machines. For this reason many people refer to TPM as "Total Productive Manufacturing" or "Total Process Management". TPM is a proactive approach that essentially aims to prevent any kind of slack before occurrence. Its motto is "zero error, zero work-related accident, and zero loss. Kodali, R., Chandra, S in his paper on “Analytical Hierarchy process for Justification of total productive management” discusses that, it is customers who actually drive our business. He discusses in the manufacturing sense, there is a need to provide the necessary production responses to satisfy and exceed expectations of the customers by adding value, quality and performance in all that we do. Company-wide TPM is about maximizing added value and eliminating waste across the supply chain in order to satisfy and exceed our customers’ expectations as depicted in the Fig. 1 [5].
2 OBJECTIVES

a. To identify the causes of failures/ Breakdown in the 4 roll Calendar Mill, 4 roll calendar Warm up mill
b. To plot the Cause & Effect Diagram for the same.
c. To carry out Why- Why analysis
d. To solve the problems using the tools of TPM.

3 EXPECTED BENEFITS OF TPM

The TPM tools are basically introduced to achieve Maximum Equipment Efficiency (MEE), Substantial increase in utilization of Machinery, reliability, decrease in the Manufacturing cost and High Quality of Products. The intangible benefits are enhanced job satisfaction and creation of safe, healthy & pleasant work environments.

4 LITERATURE SURVEY

Nackajima [1], Hartmann [2], Willmott [3] and Tsuchiya [4] identified three main objective of TPM: zero defects, zero breakdowns and zero accidents. These goals are obtained through the implementation of activities planned to increase equipment efficiency, the creation of a program autonomous maintenance, the establishing of a planned maintenance system, the organization of training course for workers and the design of a plant management system [1-4]. In the western manufacturing approach, in front of the Japanese culture, the focus of TPM is mainly on equipment, in terms of an overall effectiveness perspective [2]. In these companies TPM seeks to engender a company-wide approach towards achieving a standard of performance in manufacturing, in terms of the overall effectiveness of equipment, machines and processes, which is truly world class [3]. The value of deploying TPM is widely recognized, particularly in today’s increasingly mechanized and automated production facilities, yet it is not easy to ensure its implementation at the core of modern, large-scale production processes [4]. TPM can allow improvements in terms of equipment effectiveness, better products quality, meeting promised delivery dates and helpful work environments [5]. Companies that adopt successfully TPM are seeing 50 per cent reduction in breakdown labor rates, 70 per cent reduction in lost production, 50-90 per cent reduction in setup, and 60 per cent reduction in costs per maintenance unit [6]. Peter Willmott and Dennis McCarthy[7] in their book on TPM- A Route to World Class performance states that, the effective application of modern technology can only be achieved through people - starting with the operators and maintainers of that technology - and not through systems alone. Hence the emergence of total, productive maintenance as the enabling tool to maximize the effectiveness of our equipment by setting and maintaining the optimum relationship between people and their machines. Further [8] to [14] have highlighted the effects of TPM, TQM methods and the remedial measures to be undertaken to eliminate waste and increase productivity.
5 TOOLS OF TPM
The following modern concepts are studied in the tyre manufacturing organization. They are, Fuguai, Kaizen, OPL, GTF and Cleaning-Lubrication-Inspection, Tightening (CLIT). Fuguai is an abnormality/defect present in equipment, which needs to be rectified. Fuguai are classified as White Fuguai and Red Fuguai. White Fuguai discusses the minor problems which operators themselves can rectify. Red Fuguai are Fuguais which requires assistance from the specialized person. Fuguais are broadly classified as Minor defect, Unfulfilled basic condition, Hard to access, Contamination source, Quality defect source, Unnecessary & Non-urgent item, Unsafe place, Unsafe practice. OPL means One Point Lesson which is a pictorial representation & propagation of information to the team members. OPL are broadly classified as Basic Knowledge case, Improvement case, and Troubleshooting case. Kaizen has multiple dimensions. It means Restoration, Betterment, and Revolution & Renovation. Kaizen leads to improvement from present condition. Normally, it is an outcome of team brainstorming to reduce Time/Cost & increase Quality/Productivity or Eliminate accidents etc., CLIT is Jishu Hozen Step-III activity carried out for tentative standards. Using the CLIT symbols shown in the Fig 2, the study is carried out in various mills to improve productivity at the expense of decrease in the wastes & rejections.

6 CASE STUDY
The study was carried out in a Tyre manufacturing Industry and the following activities are undertaken. The step by step activities are as follows.
a. Drawing the Layout of the Calendar Mill – Layout of Business Unit 1 is shown in Fig. 3.

Fig 2 CLIT Symbols

Fig 3 LAYOUT OF BU1 4ROLL CALENDAR
b. Collection of Data: Fig 4 shows the various types of breakdowns & the corresponding loss in time due to the breakdowns in 4 Roll Calendar for the past one year.

<table>
<thead>
<tr>
<th>TYPES OF BREAKDOWN</th>
<th>LOSS OF TIME IN MIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MECHANICAL</td>
<td>10730</td>
</tr>
<tr>
<td>ELECTRICAL</td>
<td>2990</td>
</tr>
<tr>
<td>ELECTRONICS</td>
<td>7130</td>
</tr>
<tr>
<td>INSTRUMENTATION</td>
<td>2035</td>
</tr>
<tr>
<td>MATERIAL HANDLING</td>
<td></td>
</tr>
<tr>
<td>HOLD UP</td>
<td>6176</td>
</tr>
<tr>
<td>OPERATIONAL</td>
<td>2915</td>
</tr>
<tr>
<td>TECHNICAL</td>
<td>1595</td>
</tr>
</tbody>
</table>

**Fig 4 Types of Breakdown**

It is observed that the types of Breakdowns are high in the Mechanical sector. Hence, the study was concentrated in solving the problems in the Mechanical sector.

c. Brainstorming: The brainstorming was carried out involving the Engineers, Workers in the shop floor. The reasons for the breakdowns are discussed. The reasons for the breakdowns are tabulated. They are, overhead Conveyor failure, Breaker pad failure, Safety rope failure, Cutter failure, Lubrication failure, Motor failure, Connecting gear failure, Bull gear failure, Cooling spray pipes clogging, Journal bearings failure, Mill tray failure. Fig 5, shows the breakdown & the loss of time due to the breakdown. It is observed that failure due to Mechanical Problems is paramount.

d. Cause & Effect Diagram: A detailed Cause & Effect Diagram was plotted to identify the causes of failure & the remedial measures to be undertaken to solve the problem. Fig 6 shows the cause & effect diagram for the Warm up Mill.
e. Based on the observations, a Why-Why analysis was carried out for the Warm up Mill. Fig 7 shows the analysis for the Warm up mill.

<table>
<thead>
<tr>
<th>BU No.:1</th>
<th>Circle No.:7</th>
<th>Circle Name:</th>
<th>Why-Why No.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment Name:</td>
<td>4 ROLL CALENDER WARM-UP MILL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Break Down Actual Observation:
(Frequent failure of overhead conveyor)

What is Your Action to Correct the Problem:
(Modifying the telescopic rod assembly)

<table>
<thead>
<tr>
<th>ITEM</th>
<th>WHY</th>
<th>WHY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conveyor motor failure</strong></td>
<td>1. Overloading of the compound on the conveyor</td>
<td>1. No Proper operational awareness and attitude</td>
</tr>
<tr>
<td></td>
<td>2. Improper maintenance</td>
<td>2. Not following the maintenance schedule</td>
</tr>
<tr>
<td><strong>Conveyor drive chain slippage</strong></td>
<td>1. Improper adjustment</td>
<td>No proper maintenance</td>
</tr>
<tr>
<td></td>
<td>2. Sprocket worn-out</td>
<td>Not following the maintenance schedule</td>
</tr>
<tr>
<td></td>
<td>3. No lubrication</td>
<td></td>
</tr>
<tr>
<td><strong>Overloading of compound</strong></td>
<td>1. No Proper operational awareness and attitude</td>
<td>No proper training</td>
</tr>
<tr>
<td><strong>Pillow block tilting &amp; touching the conveyor roller</strong></td>
<td>Due to the slippage of telescopic assembly</td>
<td>Design not compatible</td>
</tr>
<tr>
<td><strong>Pillow block breakage</strong></td>
<td>Due to pillow block tilting &amp; touching the conveyor roller</td>
<td>Due to the slippage of telescopic</td>
</tr>
<tr>
<td><strong>Pillow block bearing failure</strong></td>
<td>Due to pillow block tilting &amp; touching the conveyor roller</td>
<td>Due to the slippage of telescopic</td>
</tr>
</tbody>
</table>
Frequent adjustment of belt tension  | Due to the slippage of telescopic  | Design not compatible
---|---|---
Conveyor belt shifting  | Due to the slippage of telescopic  | Design not compatible

**Fig 7 Why-Why analysis for the Warm Up Mill**

7 ANALYSIS
From the above ‘why why’ analysis we observed that the main reason for the failure of overhead conveyor was due to improper design of Telescopic rod; This telescopic rod was circular in section, due to heavy loading conditions and vibrations of conveyor the telescopic rod used to get slipped resulting in tilting of telescopic rod causing Damage of roller, Wearing of pillow block, Wearing of pillow block bearing, Machine stoppage, Extra setup time. Considering all the factors mentioned above there is a need for modifying the existing telescopic assembly, so the action taken was to modify the telescopic rod.

KAIZEN IDEA: A sample remedial measure for the telescopic assembly is shown in Fig. 8.1 & 8.2.

![Fig 8.1: Telescopic assembly before](image1)

![Fig 8.2: Telescopic assembly after](image2)

Further, Kaizen study was carried out and a detailed Kaizen sheet and the analysis for the 4 Roll Calendar Warm up Mill is shown in the Fig 8.3. It is observed that there is considerable reduction of downtime after the implementation of Kaizen.
Implementation of kaizen has reduced the number of occasions of adjusting the conveyor belt tension from existing 20 minutes per day to twice a week and also made the operation easy. This has reduced the man hour loss & production loss resulting in improved productivity. CLIT Graph shown in Fig 9 shows the decrease in the reduction in the Cleaning, Lubrication, and Inspection & Tightening periods in the Feed & Warm up Mill.

8 CONCLUSION
By implementing Kaizen 4 Roll calendar equipment availability for production has improved, in turn OEE (over all equipment efficiency) has raised considerably. Further, Kaizen implementation has resulted in the reduction of idle man hours and machine hours. Further, the case study helped in the Identification of Fuguais, applying Quality Control tools & techniques, foreseeing probable resistance and follow-up and corrective action to be taken up.
9 REFERENCES

Maintenance, TPM Press Inc., London
Productivity Press, Portland, Oregon.
maintenance.” Production Planning & Control, Vol. 12, No.7; 695-705.
Engineering, No.3: 6-66.
[7] Peter Willmott and Dennis McCarthy[7] in their book on TPM- A Route to World Class
performance
total quality management, International Journal of Quality & Reliability, 16(1), 72-84.
 perceptions, Team Performance Management, 5(1), 16-22.
productivity, Production and Inventory Management Journal, 33(4), 6-9.
reduction (SUR) beyond the productive maintenance (TPM), Proceedings of the IMECH E Part B
Using Enterprise Architecture to Gain Strategic Alignment for Project Management in Information Technology

Dr. Jack Elson, National University
Jeanne Stacy, Northrop Grumman
Brian Turner, Northrop Grumman
Russell F. Zimmermann, Northrop Grumman

Abstract

Companies with a strong foundation for execution have an increased advantage over companies that fail to plan and implement a solid foundation. Many of the world’s most successful companies thrive as a result of their commitment to create a solid foundation for business execution using Enterprise Architecture. Implementing Enterprise Architecture provides Project Management Offices (PMOs) and Project Managers (PMs) with the tools necessary to plan and manage projects successfully. This paper examines PMOs, and how PMs can use Enterprise Architecture to plan and select projects that align with the organizations’ strategies. It also presents the results of a survey of Project Management Professionals and members of an Enterprise Architecture Community of Practice about their familiarity and experience implementing Enterprise Architecture.

Keywords: Enterprise Architecture, Project Management, Strategic Alignment

INTRODUCTION

Organizations often attempt to use IT in ways that do not focus on the goals, objectives, and strategies that they are trying to achieve. Project selection is often done through political maneuvering and major IT budgeting decisions are often based on cost rather than the organization’s mission and goals (Pentecost, 2005). Additionally, organizations fail to deliver on IT projects. One survey shows that 74% of IT projects failed to deliver expected results and business performance relying on IT capabilities has declined because of misalignment and the complexity of systems, applications and other infrastructure. (Shpilberg, et al., 2007)

A solution to this problem is the use of Enterprise Architecture (EA), which is a standardized set of artifacts designed to model the structure and behavior of an organization. It can be a useful tool for selecting and planning projects by Information Technology (IT) Project
Management professionals. EA is used by managers in various business realms, including the IT industry, to describe the current and future states of an organization and the transition from one to the other.

Using EA will allow for strategic alignment which is the extent to which the business mission, objectives, and plans support and are supported by the IT mission, objectives, and plans, and the lack of it is one of the key reasons organizations fail to realize value from IT investments (Reich and Benbasat, 2000). Additionally, IT and its infrastructure are entwined with its business processes and activities that it supports and has been described as the nervous system of an organization,. As such, it can be either enabling or prohibitive and restrictive (Strnadl, 2006).

This paper will present research that supports the theory that the use of Enterprise Architecture as part of project management offers organizations greater strategic alignment and that PMOs and IT PMs should utilize an organization’s Enterprise Architecture because it will align their projects with their organization’s strategies.

BACKGROUND AND THESIS

Project Management

Project Management and specifically its use in IT is a necessity in today’s world of rapid change in order to meet corporate objectives and achieve strategic goals. Improving business processes is the name of the game and it therefore must be done successfully and in ways that is aligned with corporate strategy and also provides appropriate return on investment.,

However, when it comes to project management, organizations attempt to manage projects and their resources efficiently, when in reality, inefficiency results (Meredith and Mantel, 2005). Furthermore, organizations must execute projects within prescribed conditions or
constraints that define the environment or the results of the project. It falls on the project manager to ensure that the project operates within those constraints. In addition, those organizations should evaluate projects that align with their strategies. And although project management often produces artifacts during the planning of a project that are similar to those of an Enterprise Architecture, they are oftentimes relevant only to the project, rather than the enterprise as a whole (Whyte, 2003).

Project Management is accomplished through the use of the processes such as: initiating, planning, executing, controlling, and closing (Bidgoli, 2004). In many corporations these activities are done by two separate entities, a Project Management Office (PMO) and Project Managers (PM).

The PMO is the office or department responsible for establishing, maintaining, and enforcing Project Management processes, procedures, and standards. It provides services, support, and certification for Project Managers (CSUMB, 2007). Hobbs (2007) indicates that a vast array of other names are currently being used in management practices, such as Project Support Office, Project Office, PMO, Center of Excellence, and in many cases, names do not clearly differentiate PMOs. Regardless of what an organization may call it, the PMO essentially works toward the same outcomes. The PMO establishes, maintains, and enforces an organization’s Project Management processes, procedures, and standards; thus, it can identify those projects that an organization can properly manage.

Project Managers are responsible for the overall management of initiating, controlling, completing, and closing a specific project – bringing about its successful completion. The Project Management Institute is the professional society of overseeing project management. It publishes the Project Manager Body of Knowledge (PMBoK), a guide for its certification of the Project
Management Professional. According to the PMBoK, Project Management is the application of knowledge, skills, tools, and techniques to project activities to meet project requirements. Project Management is accomplished through the use of processes, such as: initiating, planning, executing, controlling, and closing (IEEE, 2004).

This paper asserts that an organization’s PMs and PMOs should use Enterprise Architecture to plan and select projects that align with the organization’s strategies. If the organization’s PMO and PMs use an Enterprise Architecture, then the outcome of their projects align with the organization’s strategic goals. The PMO, in this case, considers the organization’s Enterprise Architecture when selecting projects. The Project Manager considers the Enterprise Architecture while planning projects. If they do not use the Enterprise Architecture, their projects result in misalignment with the strategic goal of the organization and results in projects that achieve only a partial strategic outcome or none at all.

**Enterprise Architecture, Corporate Strategy, and Alignment**

Enterprise Architecture is a methodology that generates and uses a digital model of the firm. Bernard (2005) defines the Enterprise Architecture as “both a management program and a documentation method that together provides an actionable, coordinated view of an enterprise’s strategic direction, business services, information flows, and resource utilization” (p. 33).

EA was first developed by John Zachman at IBM in the 1980’s, and the Zachman Framework is considered the de-facto standard for the elements of Enterprise Architecture. The framework utilizes a matrix that depicts levels of models in an enterprise (scope, business or conceptual, system or logical, technology and detailed) along with interrogatives (what, how, where, who, when, and why) for each of the models to describe the ‘primitives’ of Enterprise Architecture. The Zachman model describes business processes, application architecture,
hardware, IT infrastructure, and technology standards in its model of primitives. In addition, strategy is shown as the basis of motivation for the Scope and Business Model (Zachman, 2007). A copy of the Zachman framework is included in Figure 1.

![The Zachman Framework](image)

**Figure 1.** The Zachman Framework (Source: Zachman 2007).

A primary purpose of EA is to make visible both the structure and the vision of the organization (Jonkers, et al., 2006a). And both internal and external motivators lead organizations to create an Enterprise Architecture.

Internally, organizations are driven by the analysis that is possible with an architecture. Trends and relationships between business and IT that are not readily apparent become clear when Enterprise Architecture artifacts and reports are created and reviewed (Ybarra, 2006).

The analysis may be used to determine which processes should be eliminated, added, or changed.
to improve overall productivity and determine the impact that changes to the IT infrastructure will have.

Externally, laws and governing bodies are may lead an organization to use EA. The Clinger-Cohen act of 1996 requires IT Architectures for all federal agencies. And the Sarbanes-Oxley Act of 2002 requires that companies implement internal IT controls and be able to provide detailed accounting data to auditors upon request. The Group of Ten (G10) countries and the European Union are now expected to comply with the Basel II capital adequacy framework made by the Basel Committee on Banking Supervision that establishes an international standard for how banks manage and measure risk. These laws and regulations have far reaching implications and it is not always clear what must be done to ensure compliance and EA may assist companies in complying.

Research shows that organizations benefit from having a full-time EA team to develop, support, and maintain the architecture. This team is comprised of IT professionals who are primarily concerned with fitting immediate business needs into a company’s longer-term vision (Ross, et al., 2006). And once established, it is necessary to keep an Enterprise Architecture up-to-date with the latest information. The architecture is a tool used to plan for change and to communicate information about those changes to stakeholders throughout the diverse domains of an organization. Without current, accurate data, the effectiveness of Enterprise Architecture will diminish (Jonkers, et al., 2006b).

The top level of Zachman Framework includes a model of the organization’s strategy. Organizational strategies define the plans used by an organization to align with and fulfill its mission or charter. Strategies include a description of the organization's goals and objectives and the means the organization plans to use to achieve those goals and objectives. Strategies cross
the breadth and depth of an organization, accounting for activities across all domains and levels, including corporate, business, and functional. A common means of developing strategy is the SWOT analysis as part of the strategic planning process which attempts to achieve a long-term sustainable advantage for the firm by analyzing both the environmental opportunities and threats, and the strengths and weaknesses of the organization (Hax, 1990).

Organizational strategy is a key component in Enterprise Architecture. EA commonly attempts to provide coherence to the expression and documentation of strategy as well as assist in its implementation. An Enterprise Architecture is a representation of an organization’s enterprise that should be understood by stakeholders and illustrates how strategies can and will be served by the organization’s activities (Veasey, 2001).

It is the layered structure of EA that therefore allows for alignment of the organization’s various aspects with its strategy. By definition, EA provides alignment with the organizational strategies (Bernard, 2003). Enterprise Architecture is “the organizing logic for business processes and IT infrastructure reflecting the integration and standardization requirements of the company’s operating model” (Ross, et al., 2006). It this alignment within the EA model that can be utilized by project management to align projects and project goals with the strategic goals of the organization. The alignment of projects to an organization’s strategies involves recommending, funding and completing projects that directly help the organization achieve one or more of their objectives.

Thesis

Having reviewed the background and key concepts, then formally the thesis of this paper is: Enterprise Architecture should be used by the PMO and PM because doing so aligns their
project with the organization’s strategies. In order to make its case, the paper proposes the following three assertions:

- Assertion 1, Project Managers determine how projects are implemented;
- Assertion 2, PMOs recommend which projects get funded; and
- Assertion 3, Enterprise Architecture supports an organization’s strategy.

With those three assertions established the thesis logically follows.

These three assertions have been partially explained and supported by the discussion above. The literature provides little if any support of the thesis. The paper will present the results of a survey on these topics which will further substantiate their validity along with additional support from the literature. The thesis is then deduced from this information and also shown to be valid. The summary and recommendations discuss the implications of this research and propose further work to extend the investigation.

**Survey Methodology**

A survey was designed to capture the opinions of Project Management and Enterprise Architecture CoP members. This online survey consists of 12 questions. The survey questions are designed to identify the respondents’ level of familiarity with Enterprise Architecture, Project Management, and the PMO, as well as determine their level of agreement with each of the assertions made in this paper. The questions also collect respondents’ opinions regarding the use and effectiveness of Enterprise Architecture in Project Management and their agreement with the thesis.

The survey provides anonymity to its respondents. The target audience is a group of 400 Project Management professionals and 158 members of an Enterprise Architecture CoP at a large IT contracting firm. The Enterprise Architecture CoP members are expected to be experienced,
with some being professionally certified. In some cases, however, the members are systems engineers or other IT professionals with an interest in Enterprise Architecture. The Project Management professionals are part of a user community within their place of work. They are employed as experienced Project Management professionals or certified in their profession. They were chosen for this survey to gain insight into their professional opinions regarding the topics we are researching and the assertions upon which this thesis is built. While all of these members are chosen for their experienced professional opinion, not every member of the Enterprise Architecture CoP is well-versed in Enterprise Architecture. Many are involved in the CoP to learn about Enterprise Architecture and how it can help their organizations.

Most of the questions were designed with Likert-type responses with five levels that were scored on a numerical scale of 0 to 4, with 0 indicating very unfavorable and 4 very favorable. The score of 2 indicated a neutral opinion. The results are tabulated and presented in the section Results of Survey.

LITERATURE REVIEW

Surprisingly, very little research is readily available that points directly to how a PMO or PM might use Enterprise Architecture to manage its IT projects. One study by Riempp points to the usefulness of Enterprise Architecture in managing application portfolios in large organizations. An application portfolio, in this case, is a collection of software and business functions that are managed by the organizations. The study suggests that well-integrated Enterprise Architectures would support management decision making and improve the quality of information (Riempp, 2007).

The remainder of the literature review focuses on three key concepts in the thesis statement and their relationship to Enterprise Architecture. The three key concepts are
alignment, PMO and PM, and projects with strategies. The sub-topics of the literature review include: Alignment and Enterprise Architecture; PMO, PM, and Enterprise Architecture; and Projects with Strategies and Enterprise Architecture.

**Alignment and Enterprise Architecture**

A review of the literature on the topic of Alignment and Enterprise Architecture yielded empirical studies and proposals of best practices to help organizations achieve their goals. Alignment in this context refers to linking a business process and IT to an organization’s strategies. Rico (2006) outlines a methodology for estimating the return on investment (ROI) of an Enterprise Architecture and reinforces the point that the purpose of Enterprise Architecture is to align an organization’s IT with its strategies. While the article describes the value in applying Project Management practices to the creation of an Enterprise Architecture, it does not delve into the use of the Enterprise Architecture by the PMO or PMs.

Esteva, et al. (2006) propose the use of a new seven-stage methodology, the Formal Technology Introduction Process (FTIP), and a new entity to the organizational structure, the Advanced Technology Group (ATG), for incorporating technologies into an enterprise. One of the objectives is to establish how well the proposed technology aligns with the Enterprise Architecture. The PMO is mentioned in the closing remarks as a group that would work with the Advanced Technology Group and provided oversight for the Formal Technology Process. Project Management is not explicitly identified, but many tasks performed by the ATG would traditionally be carried out by Project Managers. This article does not explicitly discuss how PMOs and PMs can utilize EA, but does indirectly recommend that the PMO and PMs use Enterprise Architecture to align with their organization’s strategies by including the PMO as a partner to the ATG and describing Project Management tasks throughout the FTIP.
PMO, PM, and Enterprise Architecture

A review of the literature regarding PMOs and PMs with Enterprise Architecture brought about a variety of important concepts, including Enterprise Architecture frameworks’ relevance to current Chief Information Officer (CIO) concerns, purported best practices for planning and controlling IT investments and cultural issues that impede progress in modernizing the IT infrastructure. However, there was little specific information on how PMOs and PMs should utilize EA.

Lindström, et al. (2006) conducted a study to prioritize CIOs’ top eleven concerns. Those concerns were then mapped to the theories of two widely recognized Enterprise Architecture frameworks: the Department of Defense Architecture Framework (DODAF) and the Zachman Framework. The research focuses on the concerns of CIOs and the alignment, as well as the lack of alignment, of Enterprise Architecture Frameworks to those concerns. The authors’ conclusions do not specifically address the PMO or PM, but IT Project Portfolio models are among the recommended additions to Enterprise Architecture frameworks to improvement to CIOs’ concerns.

PMOs, PMs, and Enterprise Architectures all contribute to Capital Planning and Investment Control (CPIC), a methodology for selecting, managing, and evaluating IT investments. The use of CPIC is mandated by the Clinger-Cohen Act of 1996 for all federal agencies in the United States government. Pentecost (2005) describes the three phases of CPIC and discusses the roles of the PMO and PMs in the process. The PMOs role is in project selection, control oversight and final evaluation, while the PM is mostly involved in the control phase. So while he describes how CPIC is utilized he does not explicitly state the importance of
the PMO and Project Management utilization of Enterprise Architecture and how to align projects with the organization’s strategies

**Projects with Strategies and Enterprise Architecture**

A review of the literature regarding Enterprise Architecture and IT projects involving corporate strategy yields little in terms of PMOs, PMs and utilizing EA for strategic alignment. Shupe and Behling (2006) stress the importance of managing an organization’s project portfolio to minimize costs, maximize productivity, and align activities with strategic objectives. However, their work does not describe the role that the PMO and PMs should play in portfolio management. And Van Grembergen, et al. (2003) described how an IT strategic Balanced Score Card can be used to show how an organization’s strategies are being served by IT utilizing a case study. The article does not specifically address the role Enterprise Architecture plays in implementing and using the BSC. It also does not outline how the PMO or PMs can or should use the Enterprise Architecture to align projects with the organization’s strategies.

**RESULTS OF SURVEY**

Project Management professionals and members of an Enterprise Architect Community of Practice (CoP) were surveyed. The Enterprise Architecture CoP is comprised of Enterprise Architecture professionals and managers that are involved in the organization for professional development. The overall percentage responding to the survey was 17.2%. This is broken down by group and summarized in Table 1 below.
<table>
<thead>
<tr>
<th>Survey Respondents</th>
<th>Number Surveyed</th>
<th>Number of Respondents</th>
<th>Percent Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Management Professionals</td>
<td>400</td>
<td>76</td>
<td>19.0%</td>
</tr>
<tr>
<td>Enterprise Architect CoP Members</td>
<td>158</td>
<td>20</td>
<td>7.9%</td>
</tr>
<tr>
<td>Total Both Groups</td>
<td>558</td>
<td>96</td>
<td>17.2%</td>
</tr>
</tbody>
</table>

Table 1. Survey Replies

Questions 1 and 2 determined the level of familiarity with Enterprise Architecture and PMOs. Question 1: How familiar are you with Enterprise Architecture? And Question 2: How familiar are you with Project Management?

Table 2 summarizes the results of these two questions. Additionally, more than half of the respondents (55% or 32) had more than five years of experience in a PMO, and 77% or 45 had more than five years of experience as a Project Manager. Within the respondent group from the Enterprise Architecture CoP, 90% responded that they were familiar or very familiar with Enterprise Architecture, 55% indicating a certification. Of the responses, all indicated they were either familiar or very familiar with Project Management, 5% indicating that they were Project Management certified.

<table>
<thead>
<tr>
<th></th>
<th>Project Management Professionals</th>
<th>Enterprise Architects</th>
</tr>
</thead>
<tbody>
<tr>
<td>How familiar are you with Enterprise Architecture?</td>
<td>60.6%</td>
<td>90%</td>
</tr>
<tr>
<td>How familiar are you with Project Management?</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2. Questions 1 and 2.
Assertion 1: PMOs Recommend Which Projects get Selected

Survey question 7 corresponds to assertion 1: “A PMO should be responsible for recommending which projects an organization should fund.” The results are summarized in Table 3.

<table>
<thead>
<tr>
<th>Project Management Professionals</th>
<th>Enterprise Architects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Rating</td>
<td>Agree + Strongly Agree</td>
</tr>
<tr>
<td>2.35</td>
<td>60%</td>
</tr>
</tbody>
</table>

Table 3. Response to Question 7.

60% of the Project Management population surveyed either agree or strongly agree with that statement based upon responses to question 6. These results appear below. As a whole, the Enterprise Architecture CoP was torn, as 50% answered neutral, agreed, or strongly agreed with the statement. These results are reflected in Figure 5. Comments, such as those listed below, from members of the Enterprise Architecture CoP generally indicated a lack of confidence in the PMO in either expertise or rigor:

- “Without EA, a PMO or PM will create their program in a vacuum, with high likelihood of overlap with other projects in an organization, poorly understood risks, and higher likelihood of poor execution, i.e. over budget or not on time.”
- “The PMO should certain have a majority say, but there should be other sources used to weigh the final decisions, such as strategic goals of the over enterprise/organization, portfolio management, cross-cutting synergies that could be achieved, etc.”
An additional question, “What are your reasons for using Enterprise Architecture in a Project Management Office”, many responses, including those listed below, indicate strategic results, funding and cost as well as structure and standards as important reasons to use the Enterprise Architecture:

- “Strategic alignment - decision optimization duplication reduction - legal conformance. i.e. Clinger–Cohen, Raines rules, etcetera.”
- “Consistency, repetition, and the cost savings standardization they bring. Allows the Program Manager to compare performance between two otherwise different projects. Also, assuming an organization larger than the PMO is funding the bulk of the infrastructure, you can take advantage of not having to re-invent (or pay for...) the wheel.”
- “Standardize Systems and Applications to reduce cost.”

Additional data from the literature supports assertion 1. Kendall (2003) estimates that today there are over 50,000 such organizations in the United States of America alone. Since the PMO has knowledge of its organization’s priorities and available resources, it is well suited to recommend which projects should be selected. PMOs perform several functions as part of the project selection process. They prioritize and promote projects that capitalize on the organization’s strengths (Meredith and Mantel, 2005), and also guide projects safely and minimize risk (Kendall, 2003). While the specific responsibilities of the PMO may differ between organizations, they essentially all work to improve outcomes on selected projects. The PMO will only recommend projects in which they determine the desired outcomes are possible to achieve.

The PMO is constantly looking to improve an organization’s Project Management practices. It is a team dedicated to improving the practice of Project Management in the
organization (Englund, et al., 2003). And it also provides the capability to ensure professionalism and excellence in applying widely accepted principles and preferred project management practices to each project effort (Hill, 2003).

However, ultimately, project selection is the primary function of a PMO. Any project selection methodology should be linked to the overall goals and objectives of the organization. “No project should be selected that does not, in some way, further the stated objectives of the organization” (Williams, 2002). “Many organizations today also have a project management office (PMO), which carries out project scoring and ranking directly or facilitates that process for stakeholders and/ or line management” (Brandon, 2005). The PMO is able to properly recommend projects because it carries out processes that “score” how well a project will further the stated objectives of an organization.

Assertion 2: PMs Determine how Projects are Implemented

It is evident through the results of the survey that PMs agree that a PM should determine how projects are implemented. Question 7 asks: “A project Manager should determine how the project utilizes the organizations…” The resources identified in the question—business processes, applications, hardware, IT infrastructure and technology standards—are key artifacts normally included in an organization’s EA. Table 4 summarizes these results.

Of the 76 Project Management respondents, the overall majority agree that PMs should determine how the project utilizes the parts of the organization that are referenced in Enterprise Architecture. Of the 20 Enterprise Architecture respondents, the majority also agree.
A Project Manager should determine how the project utilizes the organization’s …

<table>
<thead>
<tr>
<th>A Project Manager should determine how the project utilizes the organization’s …</th>
<th>Project Management Professionals</th>
<th>Enterprise Architects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg Rating</td>
<td>Agree + Strongly Agree</td>
</tr>
<tr>
<td>Business Processes</td>
<td>2.68</td>
<td>69%</td>
</tr>
<tr>
<td>Applications</td>
<td>2.68</td>
<td>67%</td>
</tr>
<tr>
<td>Hardware</td>
<td>2.66</td>
<td>63%</td>
</tr>
<tr>
<td>IT Infrastructure</td>
<td>2.64</td>
<td>63%</td>
</tr>
<tr>
<td>Technology Standards</td>
<td>2.58</td>
<td>65%</td>
</tr>
</tbody>
</table>

Table 4. Responsibilities of the Project Manager (Question 7).

From these results, it is interesting to note that while both groups agree, Project Management professionals agree more strongly with the statements that the PM should determine how the project utilizes business processes, applications, hardware, IT infrastructure, and technology standards.

Data from the literature also support assertion 2. Cleland (1998) indicates that the implementation of a project requires “that the total job be broken down into components (hardware, software, and services) and that these components be further broken down into assignable work packages.”

In order to illustrate the impact PMs have on how projects are implemented, it is important to understand their role. As indicated by Knutson, the PMs dictate the way in which time, cost, and performance objectives are met during the execution of a project (Knutson, 2001). This concept is further reinforced by the overview of Project Management Areas and Processes described in the PMBoK. During the Project Plan Development phase, the PMs should use other planning outputs, historical information, organizational policies, constraints, and assertions (IEEE, 2004).
Finally, PMs determine how projects are implemented with regard to an organization’s strategy. According to the IEEE, PMs must ensure projects address the strategy of the organization, the stakeholder expectations, and standards and regulations; this is done during the planning phases of the project plan. Each is reflected in the inputs of the plan through policy, constraints, and stakeholder requirements (IEEE, 2004).

The survey and literature overwhelmingly support the assertion that the PM should determine how the project gets implemented.

**Assertion 3: EA Supports an Organization’s Strategy**

The survey results indicate that PMs, as well as Enterprise Architects, agree that the Enterprise Architecture can provide better strategic alignment for IT Project Management. The results from question are summarized in Table 5.

<table>
<thead>
<tr>
<th>5. Enterprise Architecture…</th>
<th>Project Management Professionals</th>
<th>Enterprise Architects</th>
</tr>
</thead>
<tbody>
<tr>
<td>.. can provide better strategic alignment for IT project management?</td>
<td>Rating</td>
<td>Agree + Strongly Agree</td>
</tr>
<tr>
<td>3.16</td>
<td>87%</td>
<td>3.45</td>
</tr>
<tr>
<td>.. supports an organization’s strategies?</td>
<td>3.03</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table 5. Enterprise Architecture for strategy? (Question 5).

As evident in the results, 86.9% of PMs and 90.0% of Enterprise Architects surveyed indicated that the Enterprise Architecture could provide better strategic alignment for IT Project Management. In addition, 79.0% of PMs and 90.0% of Enterprise Architects indicated that the Enterprise Architecture supports an organization’s strategies. And also, many of the comments indicated how an Enterprise Architecture can support the performance goals and strategy. It is evident from these comments, and also, results and the research that Enterprise Architecture
supports an organization’s strategies. In addition to the survey results, data from the literature also supports assertion 3. Strategic alignment is an established practice in the business world today. Philip (2007) provides a comprehensive review of literature elaborating on the value of strategic alignment for successful firms and its importance in the recent past and says, “The message is thus loud and clear-strategic planning is an essential activity in today's dynamic business environment”. In addition, Gregor, et al. (2007) suggest that an “organization’s Enterprise Architecture may also be an enabler of alignment”.

Bernard (2005) stresses “one of the main purposes of the Enterprise Architecture program is to support and improve the enterprise’s strategic and business planning.” The Enterprise Architecture must identify and link the strategic goals, initiatives and measures to the remainder of the Enterprise Architecture. Bernard supports this concept with examples of Enterprise Architecture artifacts which demonstrate mappings of strategic business goals and strategic initiatives to supporting Enterprise Architecture components. He stresses the importance of presenting the architecture to key decision makers in a manner that facilitates greater understanding and analysis of the artifacts.

Rico offers a high-level description the purpose of the Zachman Framework, the underlying elements of Enterprise Architecture, “…the purpose of the [Zachman] layers is to align an organization’s strategy with its information technology” (Rico, 2006). Other literature agrees. “Even if the Enterprise Architecture did little else for the organization, it could be argued that it enables the integration and alignment of business strategy and IS/IT” (Gregor, 2007). As a result of a survey and analysis of 100 firms, six benefits of Enterprise Architecture were identified, one of those being strategic outcomes. The strategic outcomes included better
operational excellence, more customer intimacy, more product leadership, and more strategic agility (Ross, et al., 2004).

**Logical Analysis and Deduction of Assertion 4**

The final thesis statement rests on the first three assertions being proven in the literature and in the survey. The final survey question (12) asked both PMs and members of an Enterprise Architecture CoP to rate their agreement with this statement. Figure 9 show these results. More than 85% of the PMs and 85% of the Enterprise Architecture CoP members agreed with this statement. One interesting comment made by a respondent to this survey makes a good point. “… if ea is implemented correctly, the question you ask is moot. PMO’s and IT PM’s would be using it as a matter of course, in fact PMO’s and IT PM’s would be assigned in accordance with EA priorities as I understand it”.

<table>
<thead>
<tr>
<th>12. PMs and PMOs Should use EA…</th>
<th>Project Management Professionals</th>
<th>Enterprise Architecture CoP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>Agree + Strongly Agree</td>
<td>Rating</td>
</tr>
<tr>
<td>.. because it will align their projects with the organization’s strategy?</td>
<td>3.16</td>
<td>85.5%</td>
</tr>
<tr>
<td></td>
<td>3.3</td>
<td>85%</td>
</tr>
</tbody>
</table>

Table 6. PMs and PMOs Should Use EA (Question 12).

Up to now the paper has shown that PMs determine how projects are implemented, and if they utilize the business processes, data models, and organizational standards from their Enterprise Architecture, then their projects will align with their organization’s strategies. It has also shown that PMOs recommend which projects should be selected. Using the information in an Enterprise Architecture will support the decision making process to ensure the projects that they recommend align with the organization’s strategies. Finally, it has shown that an Enterprise Architecture supports an organization’s strategy. The strategy is part of the Enterprise
Architecture by definition. This result is shown in literature, as well as the survey. Each of these assertions, as well as their links back into the thesis statement, is highlighted in Table 7.

In this table, the first column shows each of the assertions, or the data, for argument analysis. The second and third columns, EA Utilization and Alignment, link or warrant the information from the assertions to the other parts of the thesis. These columns explain how the Assertions themselves are tied back into the final thesis statement; thus, proving the accuracy of the statement or argument.

<table>
<thead>
<tr>
<th>Assertion</th>
<th>EA Utilization</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Managers determine how projects are implemented.</strong></td>
<td>Enterprise Architecture answers the following questions: What Business Processes, data models, and organizational standards for activities and technologies are required? (Ross, et al., 2006)</td>
<td>“By engaging with projects early and regularly, a company can ensure that each project helps build out the architecture and that the architecture is realistic and aligned with company goals” (Ross, et al., 2006).</td>
</tr>
<tr>
<td><strong>Project Management Offices recommend which projects should get selected.</strong></td>
<td>Studies indicate that well integrated Enterprise Architectures support management decision making and improve the quality of information (Riempp, 2007).</td>
<td>A process managed by the PMO “attempts to link an organization’s projects directly to the goals and strategies of that organization” (Meredith and Mantel, 2005). A PMO should “Determine if a new project is a good “fit” for the changing organization” (Meredith and Mantel., 2005).</td>
</tr>
</tbody>
</table>
Enterprise Architecture supports an organization’s strategy.

N/A

“The purpose of the [Zachman] layers is to align an organization’s strategy with its IT” (Rico, 2006).

By definition, Enterprise Architecture provides alignment with the organizational strategies (Bernard, 2003).

Table 7. Summary of Assertions Linked to Thesis.

Summarizing the information in the table, the first row demonstrates how the Project Manager utilizes Enterprise Architecture to answer relevant Zachman questions when planning projects. It also shows how the Project Manager using Enterprise Architecture supports the alignment. It continues by indicating that the Project Management Offices utilizes Enterprise Architecture by integrating it into the decision making process and ensuring alignment with the strategy. Finally, the table presents the evidence Enterprise Architecture supports an organization’s strategy by providing the alignment. The conclusion logically follows that the Project Management Office and Project Managers should use Enterprise Architecture because it will align their projects with the organization’s strategies.

CONCLUSIONS AND RECOMMENDATIONS

From the results of the literature research and the survey of Project Management and Enterprise Architects, we conclude that PMOs and PMs should utilize EA because it will align their projects with the organization’s strategies. And as a result, an organization more effectively supports strategic outcomes by utilizing Enterprise Architecture in its project selection and planning processes.
To the Project Manager, the results of this thesis mean that the Enterprise Architecture is an important factor to consider while planning projects. PMs need to incorporate standards and regulations in their projects. These are both elements of Enterprise Architecture and are incorporated by Zachman as the ‘primitives’ of the Enterprise Architecture. In addition, the results of the survey indicate that the majority of both PMs and Enterprise Architects agree that ‘The Project Manager should determine how the project utilizes the organization’s applications, business processes, hardware and IT infrastructure.’ (Question 7) While both groups agree, Project Management professionals more strongly agree with the statements. This might be for the simple reason that Project Management as a profession is more mature than Enterprise Architecture. There are simply a greater number of PMs, and the experience levels are higher in this survey population.

While many comments in the survey indicate how the Enterprise Architecture is useful to the organization implementing projects, one comment succinctly summarizes the idea: “Proper enterprise frameworks consist of relevant structure within the organization including business, applications, technology and data. This framework will provide a taxonomy and ontology that clearly identifies what processes a business performs and detailed information about how those processes are executed. The end product is a set of artifacts that describe in varying degrees of detail exactly what and how a business operates and what resources are required.”

Since some of the responses to the survey indicate that both PMs and EAs do not think that PMOs and PMs should utilize EA, this indicates the need for increased communication and education of field professionals regarding the elements of Enterprise Architecture. Future research might ask the question ‘how can Enterprise Architecture be more fully incorporated into
It would seem to be most effective to do this type of study on a case by case basis and using real world examples documented in case studies.

To the PMO that is responsible for selecting projects and guiding the organization toward achieving its strategic outcomes, the results clearly point to the value of Enterprise Architecture as it supports organizations’ strategies. The surveys of professionals’ results agree with this conclusion. In addition, answers to the question “What are your reasons for using Enterprise Architecture in a Project Management Office?” repeatedly point to standardization and funding or cost, as well as strategy. Some comments allude to the maturity of the organization’s PMO. While this aspect is beyond the scope of this paper, research into the correlation between the Project Management maturity and Enterprise Architecture utilization would further clarify the value of the creation and use of Enterprise Architecture.

In this case, a future question may be: What areas of the Enterprise Architecture might be enhanced by the project to more fully support future projects? It is evident from the survey results that increased communication between Enterprise Architects, PMs, and the PMO is necessary. One interesting response to the survey summarizes that effective utilization and maintenance are two main issues facing PMs, PMOs, and others using Enterprise Architecture within the project selection or management process. The respondent said, “the challenge is not getting to an EA, the challenge is how to make it useful and actionable. Many times, an EA may exist, with clear implications and possible impacts; but, there can be no enforcement or incentives that keep it relevant - and it withers quickly”. There is clearly a need for further work on the part of all professionals to work jointly toward the formal incorporation of Enterprise Architecture into these processes.
To the Enterprise Architect, these results indicate the need to consider Project management in developing the Enterprise Architecture. Overall, the survey results indicate a lack of awareness of Project Management on the part of Enterprise Architect professionals in the numeric results as well as in the comments. The Enterprise Architect might attempt to bridge this gap by designing tools and products which are more available or accessible to PMs.

It might also require Enterprise Architects to consider the needs of PMs (e.g., reporting needs such as risk management and communications management plans as mentioned in the PMBoK). Future study might survey Enterprise Architects to determine how the Project Management inputs might be better served by the architecture.

In all cases, further research is required. It is evident from comments made in the surveys from Project Management and Enterprise Architecture CoP members that incorporating Enterprise Architecture must be made on a case-by-case basis, depending on the accuracy and relevance of the Enterprise Architecture. As noted by Shpilberg, et al. (2007) “The complexity doesn’t magically disappear just because an IT organization learns to focus on aligned projects rather than less aligned ones.” Survey respondents pointed out that Enterprise Architectures must be updated and accurate. This subject would benefit from extensive case study reviews of Enterprise Architectures used by managers within projects and programs.

The final recommendation is that further research be conducted by the Project Management Institute (PMI) or within organizations to increase the level of communication between the Project Management and Enterprise Architecture CoP members. The Project Management professional certification should include Enterprise Architecture within its educational literature to increase awareness of the possibilities of its use.
REFERENCES


