THE FIRST
CS3 PhD SYMPOSIUM

Feb 26 2010
Message from General Chair

Swinburne University's Faculty of ICT is committed to scientific discovery and technological innovation through excellence in research. The scope of our research spans a broad spectrum - from fundamental discoveries in astronomy, through technical innovation in computing and networks, to the management of the application of technology in organisations.

The Faculty of Information and Communication Technologies at Swinburne is host to four of the key University research centres, one of which is the Centre for Complex Software Systems and Services (CS3). CS3’s mission is to significantly advance the theory and practice of complex software systems and services, and deliver high quality research results and leading-edge technology solutions of national and international benefit to business and community.

The Centre carries out research across Component Software Technology, Intelligent Agent Technology, Web and Data Technology and Workflow Technology, targeting Service Oriented Systems, Enterprise Software Systems, Social Software Systems, and Cloud Computing Systems. The Centre provides a supportive and vibrant research environment for research staff and students. CS3 has about 30 PhD and other research students, pursuing research across a diverse range of areas within the Centre.

These proceedings are of the first annual CS3 PhD Symposium. Each PhD candidate has written a three-page abstract of their thesis, and presented their work in front of peers at the PhD Symposium. The 2010 CS3 PhD symposium gave CS3 students a friendly and collaborative environment in which to present and receive feedback on their research topic. Students were given the opportunity to communicate their research ideas and receive comments from other students and staff of diverse backgrounds and fields of interest. Submissions focused on each student’s research topic for candidature, and included important information such as the problem their research attempts to solve, their approach and methods of validation. Topics were as diverse as agent-based negotiation, control theory, and software-based intelligent decision support systems.

Each submission was peer-reviewed by three other CS3 PhD students, providing detailed and valuable feedback to students, in addition to providing them with reviewing experience.

The PhD Symposium ran over a full day in February 2010, and gave each student the opportunity to present their work to their peers and FICT staff members. Students received valuable feedback on both their research topics and the technical merit of their presentations.

The quality of abstract submissions and presentations was high, and we look forward to another CS3 PhD symposium in 2011.

(General Chair)

February, 2010
The submissions for the first CS3 PhD Symposium reflected the breadth and depth of research that is being carried out in Center for Complex Software Systems and Services (CS3), at Swinburne University. The idea was to provide an opportunity to CS3 students to share their ideas with a broader community as well as to be acquainted with the research carried out by their colleagues. Current CS3 students, who have started the PhD candidature on or before 25th of July 2009, were eligible to submit their research proposals to the CS3 PhD Symposium. Same set of student were selected for the reviewing committee, in order to give them an opportunity to gain experience in the reviewing process. There was total number of 24 submissions. Each reviewer received three submissions to review. Accordingly, each submission is reviewed by three reviewers. Papers were evaluated based on use of language, how well were the arguments supported, technical soundness of the proposed solution, paper structuring and clarity in explanations. The final average score is calculated based on all the review scores. We value the enthusiasm of participants both in writing and evaluating papers amidst their usual studies and other activities. This volume contains all the revised papers. We hope that the papers are interesting and vitalizing.

(PC Chair)

February, 2010
Table of Contents

1. Minyi Li  
   On Efficient Mediation Approach to Multi-issue Negotiation with Optimal and Fair Outcomes  
   1-3
2. Garth Heward  
   Dynamic, Performance-Aware Monitoring and Management for Service-Oriented Software Systems  
   4-6
3. Tharindu Patikirikorala  
   Multi-Model Switching and Tuning Adaptive Control for QoS management in software systems  
   7-9
4. Syed Galib  
   Adaptive Traffic Assignment with Self-organizing Multi-Agent System  
   10-12
5. Sira Yongchareon  
   User Interface Model for Artifact-Centric Business Processes  
   13-15
6. Khalid Mansour  
   On-the-Fly Coordination of Automated Negotiations  
   16-18
7. Cameron Hine  
   Large-Scale Emulation of Enterprise Systems  
   19-21
8. Wojciech Lorkiewicz  
   The Resolution of Symbol Meaning in Multi-agent Systems  
   22-24
9. Alvaro Monsalve  
   Lifetime Optimization of Wireless Sensor Networks  
   25-27
10. Tan Phan  
    Specification, Integration and Management of Security in Service-Oriented Software Systems  
    28-30
11. Grzegorz Popek  
    Multi-agent System for Local Data Propagation in City Traffic Monitoring: Integration of Modal and Fuzzy Methods for Agent’s Knowledge Representation  
    31-33
12. Malinda Kapuruge  
    SeRenDiP: Towards Service Relationship Driven Processes  
    34-36
13. Rui Zhou  
    Answering XPath Queries Using XPath Views  
    37-39
14. Muhammad Ashad Kabir  
    CORE: A Framework for Context-Aware Interactive Systems with Comprehensive REasoning Technique  
    40-42
15. Indika Meedeniya  
    An Incremental Methodology for Quantitative Software Architecture Evaluation with Probabilistic Models  
    43-45
16. Tuan Nguyen  
    A Feature-Oriented Approach for Web Service Customization  
    46-48
17. Xiao Liu  
    Key Research Issues in Scientific Workflow Temporal Verification  
    49-51
18. Tran Trong Hieu  
    Integration Method for Policy Integration  
    52-54
19. Hai Huang  
    Selectivity Estimation for SPARQL Graph Pattern  
    55-57
20. Dong Yuan  
    Data Management in Scientific Cloud Workflow Systems  
    58-61
21. Jiajie Xu  
    Resource Planning for Business Process Systems  
    62-64
22. Aldeida Aleti  
    Parameter control in evolutionary algorithms  
    65-67
23. John Moore  
    Towards a more accurate Census: Integrated Intelligent Decision Support for Field Design and Management of Census Operations in Australia  
    68-70
24. Jan Richter  
    Multistage Fuzzy Decision Making in Bilateral Agent Negotiation  
    71-73
On Efficient Mediation Approach to Multi-issue Negotiation with Optimal and Fair Outcomes

Minyi Li
ICT Faculty
University of Swinburne
Hawthorn, Australia, VIC 3122
myli@swin.edu.au

Abstract—Empirical evidence suggests that self-interested agents often fail to reach optimal agreements in multi-issue negotiations. Unfortunately, most existing works for increasing the optimality of the negotiated agreements either do not address the fairness issues; or ignore the computational concerns. To address these problems, the aim of this research is to introduce an efficient mediated negotiation approach to support multiple agents reaching an optimal and fair agreement under incomplete information. We use a trusted, non-biased mediator to coordinate the negotiating agents, while also protecting the negotiating agents from unnecessary disclosure of information to their opponents. We separate out the negotiation problems over continuous issues from those over discrete issues; and investigate different mediation techniques to deal with the negotiation problems over different type of issues respectively.

Keywords—multi-agent negotiation; fairness; Pareto efficiency

I. INTRODUCTION

Negotiation is a fundamental interaction mechanism in multi-agent systems. It enables self-interested agents to act cooperatively and benefit from mutually preferred agreements [2], [3], [5], [8], [9]. Such negotiations can involve, for instance, negotiation over the package deals in agent-based trading systems, negotiation about the resource allocation among different interest groups or departments, or negotiation between the service providers and customers to reach service level agreements as to enhance the quality of services, etc. [2], [5], [11]. When multiple issues are involved in negotiation simultaneously, like price, quality attributes, delivery time, etc., the agents with divergent preferences may achieve better agreements on issues that are most important for them by trading off some on those not so important [5], [9], [12]. Such situations where all the parties are better off, are normally called “win-win” situations [5], [9], [12].

However, empirical evidence suggests that self-interested agents often end up with inefficient results in multi-issue negotiations, even though a compromise does exist that the agents could have made others they all would have preferred [7], [3], [9], [12]. Lax and Sebenius [7] discuss the Negotiator’s Dilemma in deciding whether to pursue a cooperative or a competitive strategy at a particular time during negotiation. Fatima et al. [3] point out that self-interested agents would like to reach an agreement that is as favorable to them as possible, whereas the final decision is jointly made and need to be agreed to by both the agents. Consequently, the problems met by the negotiation agents are not only to choose cooperative or competitive strategies, but also to consider how much they could gain individually if they cooperate and in which way of cooperation they could gain more, or at least receive a fair deal. Negotiation therefore, requires techniques that deal with rational agents fairly and lead them to mutually beneficial agreements. In classical negotiation theory, the typical solution proposed is the use of an independent mediator, which generally assumes that the perfect information of the negotiation parties is available for the mediator to compute the optimal outcomes and the computational concerns are often ignored.

As to address the above issues, the aim of this research is to introduce an efficient mediated negotiation approach to support multiple agents reaching an optimal and fair agreement over multiple issues under incomplete information. We use a trusted, non-bias mediator to coordinate the negotiating agents, while also protecting the negotiating agents from unnecessary disclosure of information to their opponents. We separate out the negotiation problems over continuous issues from those over discrete issues (the issues that take a finite set of values), investigate different preference representation model for different type of issues, and propose the according mediation techniques to deal with the negotiation problems with different type of preferences respectively.

The remainder of this proposal is as follows. In Section 2, we review some of the related literatures. Section 3 analyses the research problems and presents our methodologies on the designing the negotiation framework for continuous issues and discrete issues respectively. Finally, Section 4 outlines the future work in the upcoming one and a half years.

II. REVIEW OF THE LITERATURE

1) Multi-issue negotiation with utility functions: Most of the existing works have been dealing with the utility-based negotiation problems, where the agents’ preferences
are mathematically represented by utility functions. For instance, Fatima et al. [3] propose an agenda-based framework for multi-issue negotiation under time constraints in an incomplete information setting. While the authors assume that, the utility functions of the agents are linear additive. Ehtamo et al. [2] present a mediation-based gradient search method for making trade-offs, while also creating joint utility gains for the negotiating agents. However, their proposed approach leaves the fairness issue between the agents’ utility gains largely unanswered. Another mediation-based negotiation model with incomplete information is given by Lai et al. [5]. In their approach, the mediator conducts a Pareto efficient enhancement for a proposal in each negotiation period. The algorithm they develop is of high efficiency in the two-issue cases, however, it is not necessarily feasible and it does not guarantee Pareto optimality.

2) Qualitative preference and collective decision making: Utility functions are a powerful form of knowledge representation. Unfortunately, in many situations, the utility-based preference elicitation is complicated and typical users may not be able to provide much more than qualitative rankings of outcomes [1]. The researchers in AI have been developing languages for representing qualitative preferences in a succinct way, exploiting structural properties such as conditional preferential independence. Boutilier et al. [1] introduce a qualitative, graphical model of preferences, called CP-net (Conditional preference networks), which specifies individual preference relations in a relatively compact, intuitive, and structured manner.

Most existing works on CP-net focus on individual preference reasoning, including outcome optimization and comparison (See [1]), while negotiation involves multiple agents and the agents’ preferences are not common knowledge. Rossi et al. [10] define a multi-agent extension to CP-nets and propose various voting semantics for aggregating multiple agents’ preferences which are represented by CP-nets. However, they do not address computational issues. Lang [6] reconsider voting and aggregation rules in the case where the agents’ preferences have a common preferential independence structure. The author addresses the decompositions of a voting rule following a linear order over variables. However, sharing common preferential independencies over all the agents is a demanding assumption that is unlikely to be met in practice. Furthermore, the above methods assume having complete information about each agent’s CP-net, which is particularly hard to be applicable in the real world scenario.

III. RESEARCH PROBLEMS AND METHODOLOGY

A. Research Problems Summarize

To summarize, multi-issue negotiation is complex and challenging because of the following reasons.

- Incomplete information. In a multi-issue negotiation, the preference of an agent over multiple issues is complex and the outcome space is m-dimensional (m > 1) rather than a single-dimension line as in a single-issue negotiation. This makes the negotiation strategy in multi-issue negotiations complex. Increasing efficiency and fairness in multi-issue negotiation therefore, requires agents to share preference information, while disclosure of an agent’s preference to the opponents puts it at a disadvantage in a negotiation. Consequently, in most of the real world applications, the agents’ preferences are not common knowledge. Under incomplete information, the burden of computation and reasoning for the negotiation strategy become even higher, and thus it is difficult to reach efficient and fair outcomes.

- Computational complexity. In both individual and collective decision making with qualitative preference, the space of possible outcomes from which the agent (or the group of negotiating agents) has to choose often has a combinatorial structure (The number of all possible outcome is exponential in the number of variables). Much work in this field has concentrated on normative questions and on establishing abstract results regarding the possibility of designing mechanisms meeting certain requirements. Computational concerns, however, have mostly been neglected. For instance, what is the computational complexity of the mechanisms? What are the appropriate algorithmic techniques for these problems? What will happen if the number of possible outcomes to choose from becomes very large?

B. Methodology

To address the above issues, the aim of this research is to introduce an efficient mediated negotiation approach to support multiple agents reaching an efficient and fair agreement under incomplete information. We use a trusted, non-bias mediator to coordinate the negotiating agents, while also protecting the negotiating agents from unnecessary disclosure of information to their opponents. We separate out the negotiation problems over continuous issues from those over discrete issues, investigate different preference representation models for different type of issues, analyse the system goals of the negotiation problem with different type of preferences (e.g. Pareto efficiency, fairness, computational efficiency, etc.), and propose the according mediation techniques to deal with the negotiation problems with different type of issues respectively.

For the type of continuous issues, we consider the classical negotiation theory that mathematically represents agents’ preferences by utility functions. We propose a new mediated negotiation approach to support the negotiating agents reaching a Pareto optimal and fair agreement over multiple continuous issues under incomplete information. The proposed approach uses a non-bias mediator as a tool for step-by-step creation of fair joint gains. At each stage of negotiation, the mediator searches for the compromise direction based
on the solution to a mathematical programming problem, called the DMP (Deviation Minimization Problem). The objective of this approach is to find more efficient outcomes, which improve all the agents’ utilities while minimizing the difference between the agents’ utility gains, leading to fair agreements. We conduct a large amount of experiments and analyse the experimental results of the proposed approach in the context of several well-known social welfare metrics. The experimental results demonstrate that the proposed approach not only guarantees Pareto optimality, but also produces the outcomes that are close to the fair Egalitarian solution.

For the type of discrete issues, we investigate the theory of CP-net (Conditional Preference Network) as a formal model for representing and reasoning with the negotiation agents’ preference. There are not much works for aggregating multiple agents’ preferences represented by CP-nets. While the existing works either do not address computational issues, or depend on a strong assumption that all the agents share a common preferential independency structure. We introduce an efficient mediated negotiation approach for negotiation with CP-nets under incomplete information, which also allows the agents to have different preferential independency structure. The proposed approach involves a recursive procedure, called MNCP, to generate a small fair set of Pareto optimal outcomes, which is the basis for choosing the final outcome preferred by multiple agents. We prove that the candidate outcomes generated by MNCP are guarantee to be Pareto-optimal. We also conduct a large number of experiments with different scenarios varying agents’ preferences and the number of variables. The experimental results demonstrate that the propose approach is computationally efficient and produces the results that are close to the corresponding social welfare metrics.

Published Paper


IV. Future Work Plan

During the next one year (2009.10 - 2010.10) of my Phd, the main focus is on mediated negotiation with structural preference (i.e. CP-nets and its extensions). In the previous work, we have not explored more powerful variants such as TCP-nets for representing agents’ preferences in negotiation, although they can be similarly applied to support more expressive preferential semantics on preference trade-offs such as relative importance and conditional relative importance. We plan to explore in more detail how best CP-nets can improve the representation of the agents’ preferences in negotiation and produces better outcomes for multiple agents. In addition, future work also includes the exploration of possible ways for generating the optimal outcomes in different semantics including Max, Majority and Lex optimal. During the final six months (2010.10 - 2011.04), the PhD thesis which enhances and summaries our entire work will be completed.

REFERENCES

Dynamic, Performance-Aware Monitoring and Management for Service-Oriented Software Systems

Garth Heward
Faculty of Information & Communication Technologies
Swinburne University of Technology
P.O. Box 218, Hawthorn, VIC 3122, AUSTRALIA
gheward@swin.edu.au

Abstract—Monitoring of web services at run-time is required in order to ensure functional correctness, as well as measure the quality of service provided by web services. However, monitoring has an impact on the systems that it monitors and it is this impact that my research intends to measure and minimise. This optimisation will maximise value to web service providers. Additionally, web services system aspects may be managed at run-time in order to maximise value for web service providers by maximising quality of service provided. My research intends to develop a method to perform this optimisation at run-time, and simulate the effects in order to measure the optimisation’s effectiveness.

I. INTRODUCTION OF PROBLEM

The growing Service-Oriented Architecture paradigm provides distributed software with loose coupling, user-level composition, and a high level of business support through aspects such as service level agreements and various management standards [1]. Using a Service-Oriented Architecture, businesses define services, which are high-level software artefacts that directly represent business services. Web services represent the dominant means for implementation of a Service-Oriented Architecture. A set of XML standards define web services technology.

Unlike traditional software, consumers discover, bind to, and invoke web services over the Internet during run-time. Service providers can compose (assemble) web services at run-time. Since these compositions of web services are dynamically created at run-time, their properties (such as response time and correctness) are usually only available at run-time via monitoring and/or bind-time testing. Thus, monitoring web services increases confidence that services are meeting requirements at run time [2]. Monitoring of web services is required in order for service consumers and service providers to measure various quality aspects of the web service system. Service providers, web services middleware providers, service consumers and other parties may be interested in both the functional and non-functional properties of web services and their supporting infrastructure. Service providers use the results of these monitoring mechanisms for the management of the web services and their supporting infrastructure. Web services management provides the possibility to repair faults in a web service or supporting component in order to prevent errors, or modify the web service’s supporting infrastructure in order to return a higher quality of service to web service consumers.

For this research, monitoring refers to the measurement and initial analysis of information (e.g., measuring the response time of a service invocation and determining if it is within acceptable levels). Management refers to any actions on the web services system, such as replacing services within a composition, re-composing a composition, or administration of underlying infrastructure such as routers and servers.

There are numerous non-functional properties that one may wish to monitor and manage in a service-oriented system, such as performance (response time, resources consumed, throughput), security (security model, trust in partners, certificate quality, and key quality), reliability, and availability [3], [4]. Various parties involved in using or providing web services are interested in various quality aspects. A provider of an advertiser-supported news portal based on web services may be interested in the response time of their services, and a provider of web 2.0 video streaming may be interested in availability and security. In these cases, run-time monitoring and management would increase confidence in the quality of the software, and allow for increased quality of software due to an increase in the number of faults that are repaired after being detected. It is conceivable that even in these simple scenarios with only one or two quality aspects being monitored, the expense of monitoring the entire system all of the time may be greater than the benefit provided by doing so [2].

A solution that allows parties interested in monitoring the system to select exactly what aspects are monitored, at what resolution, and for what qualities, would allow users of monitoring services to obtain better value from the monitoring system. Thorough literature reviews have not discovered any systems with the desired properties (dynamic and adaptive monitoring, and monitoring trade-off) described above. My PhD investigates whether efficiency can be gained via careful selection of the details of monitors, or by making the monitoring system adaptive, so that it changes its behaviour based on the behaviour of the system being monitored. This research then intends to investigate whether a more efficient monitoring system will allow for more aspects/qualities to be monitored, or for a lower total cost of monitoring.

Monitoring a system provides information on its recent state. In order to improve the system, this information should be used for management of the system. Therefore, this research
will include a web services management system that will complement the monitoring system. The management system will use the output from the monitoring system to manage the web services and their supporting infrastructure at run-time. Such a management system would automatically/semi-automatically modify aspects of the system such as services within a composition, entire compositions, or supporting infrastructure, in order to prevent failures or increase provided quality of service in order to meet system requirements.

II. RELATED WORK

There exist research efforts for monitoring and managing the functional and non-functional properties of services and service compositions. Below is a selection of papers from a literature review on service monitoring, and a literature review on service management. These papers are provided because they are the most relevant to the research described here, in that the papers cover dynamic monitoring, pro-active management, and adaptive management.

[2] presents a method for dynamic monitoring of BPEL processes, which extends previous work [5] by adding QoS related monitoring rules. The authors list the set of unsolved problems in BPEL dynamic monitoring as allowing the application designer to select the QoS values of interest for monitoring, select how those QoS values are measured/gathered, change service monitors at run time, and change the level of monitoring at run time in order to balance monitoring with performance. These are similar problems to those defined in this proposal.

The approach uses Monitoring Rules, external to BPEL, which are used to control each BPEL process. Each Monitoring Rule can be enforced on different BPEL processes, allowing reuse of the rules. The Monitoring Rules are weaved into a BPEL2 file at deployment time. The rules are created with an equivalent of debug levels (1-5), which allows for performance versus monitoring trade-offs at run-time. Although the method allows for performance versus monitoring trade-off, the implementation is a basic five levels of monitoring. Having the ability to set that importance on not only monitors, but also services and qualities of service would improve this. For example, instead of a user giving ‘Monitor A’ precedence of level five, the user could give ‘Response Time’, or ‘Service A’ a precedence of level five. This would allow the user to concentrate the monitoring efforts more flexibly. Having a more detailed importance function than levels from one to five (e.g. ranking) would also allow a greater level of control over a larger system.

[6] presents a web services management approach designed to monitor and dynamically reallocate local system resources for services based on comparing current quality of service to Service Level Agreements. The method takes the performance impact of reallocating resources into account. The system translates high-level QoS descriptions (e.g. “Complete processing by 13:00”) to low-level resource requirements (e.g. dedicate five specific nodes to the task). These high-level requirements can be reused, for example if the requester is submitting the same job each week.

Service consumers provide the relationship between how much they are willing to pay and desired response time for each service. The performance impact is considered when reallocating services and takes into account impacts such as reloading software or libraries, moving services to different nodes, or reconfiguring networks, clusters, or file shares.

The solution allows for proactive management of the system to achieve greater benefit to service consumers at run-time, by re-allocating resources based on current loads and requests. However, the solution optimises the total quality of service/cost trade-off for individual services only.

[7] presents an architecture (Cremona) for dynamic creation and monitoring of WS-Agreement Service Level Agreements for web services. Cremona is a middleware that provides implementations for the required WS-Agreement interfaces, management for agreements, and service environment abstractions to simplify implementation. The architecture is aimed at customised QoS guarantees. The aim is to have Service Level Agreements that define the QoS guarantee for particular times and usage patterns.

When the service provider receives a request, it determines if it can meet that request. A monitoring agent in the system detects service agreement violations and predicts future violations based on these. Service reallocation may occur in these circumstances.

[8] presents a middleware framework for management and monitoring of web services. The authors aim to have a management system that can identify the source of any SLA violations in a business process on a web services system. The authors have identified the need for the management system to be efficient enough to manage and monitor large numbers of services without affecting the execution of the business process. The framework calls for service providers to publish or allow access to the internal state of their services. Service providers select the amount of information that is available for their service. The service provider may provide the Accountability Authority with logs of CPU, memory, and network data, or real-time monitoring of services by agents.

The system provides an ‘Accountability Console’, which provides users with the ability to register business processes, configure parameters of an Accountability Authority, and view diagnoses of the current web service system. The console is based on WSDM.

The authors have extended BPEL to incorporate QoS requirements, resulting in BPELQ. This is used by a service consumer to create business processes with QoS constraints across the entire process via a QoS-enabled service broker.

When a fault is detected at run-time, the system performs a root cause diagnostic to determine the most likely cause, which involves polling possibly faulty services. Once the faulty service has been identified, the system generates a new, equivalent process without the faulty service. Repair involves selecting a replacement service based on the fault detected - e.g. if a fault is detected in the network of a service, then a replacement service from the same network will not be used. The system has been prototyped as a Mule ESB extension.
The performance impact determined was between 1-7% for local agents, and negligible for the service bus (since it is just a passive listener).

In summary, there are various solutions to monitoring and managing web services systems, however only a few of these solutions cater for advanced techniques such as dynamic monitoring, adaptive monitoring, automated/assisted management, and performance trade-offs at run-time. No solution uses more than one or two of these techniques, and no solution uses techniques such as dynamic or adaptive monitoring in the manner described in this research proposal. There is no combination of these systems that meets the goals of this research. Although some of the works from the literature review provide performance measures [6], [8], there is no real comparison between techniques and no in-depth discussion of the performance impact of monitoring and management in general.

III. RESEARCH CONTRIBUTION

As shown, there exist gaps in current web service monitoring and management systems that if filled, would allow for: The ability to monitor web services with a known, and reduced performance impact; The ability to more efficiently analyse monitoring results and manage a monitored system at runtime in order to optimise it or prevent errors; Greater coverage of a monitored system, due to higher efficiency of monitors; The ability to have monitors automatically modify their behaviour based on monitoring results; The ability to estimate the response time of an invocation of a service composition based on historical information for service invocations and the current input set; And The ability to take automated or semi-automated management actions based on the output of the monitoring system.

These concepts may be generalised for the monitoring and management of any component-based software system. The research is grounded in a theoretical framework. The framework contains:

- Models of what is to be monitored (web service, business service, service host, service provider, etc.); Models of monitors themselves, as well as the monitor and management framework - information flow; Models for analysing qualities of service (either generic for plugins or specific QoS aspects); and
- An analysis and model of the impact of monitoring on the performance of the monitored system.

IV. PROJECT STATUS AND FUTURE WORK

As stated, I have performed literature reviews in the fields of web service monitoring and management in order to determine the state of the art in dynamic monitoring and management of web services. In order to verify that optimisation of web service monitoring systems is beneficial, the original unoptimised costs of monitoring need to be known. Since no existing work measured the costs of web service monitoring, I developed a classification of monitoring techniques, and setup a system to comparatively benchmark each technique. The details of this work are being published in [9]. Having established that different monitoring techniques and load levels have significantly different degrees of impact on the quality of web services, I developed a system to simulate the effect of optimisation of web service monitors. A thorough benchmark has been performed establishing the cost of each type of monitor (eavesdropping, proxy, and probe). These costs, in addition to randomly generated utility and load functions, will be provided to an optimiser.

The web service optimiser (currently incomplete) will be developed as both an enumerative (brute force) and heuristic optimisation routine. Since the search space for the problem is too large for runtime brute force calculations, the brute force algorithm will only be used in order to determine the effectiveness of any heuristic algorithms used. Each optimiser will determine the currently optimal monitoring configuration, in terms of which services to monitor, at what level (percentage of messages analysed), for what qualities, and using which monitors. This optimal configuration will then be used to re-test the system in order to measure any benefit gained.

After the completion of the web service monitoring optimisation, the same general techniques will be applied to web service management. This will provide the ability to optimally manage web services at run-time using techniques such as dynamic service load allocation and dynamic service selection. For example, if the current system load is very high, an optimal choice may be to offload service requests to a third-party service provider and pay any related expenses, rather than perform requests locally and decrease quality of service to below an acceptable level.

REFERENCES

Multi-Model Switching and Tuning Adaptive Control for QoS management in software systems

Tharindu Nishad Patikirikorala
Centre for Complex Software Systems and Services
Swinburne University of Technology
John Street, Hawthorn,
Victoria, Australia 3122
e-mail: tpatikirikorala@groupwise.swin.edu.au

Abstract—Control theory is a promising approach to provide QoS assurance in increasingly complex service based computer systems. A number of control loop schemes have been proposed for computer systems each with advantages and disadvantages. We provide a critical review of these approaches to assess their suitability for providing QoS assurance. Different operating regions of the software systems and dynamic, unpredictable environmental changes, demands fast, accurate and stable control techniques. To this end we propose the application of Multi Model Switching and Tuning adaptive control scheme to this domain and show how it can potentially address some of the shortcomings of earlier approaches. We demonstrate the feasibility of this proposal by critically evaluating the behaviors of software system and using a model problem.

Keywords-Quality of Service; Multi Model; Adaptive control

I. INTRODUCTION

Providing satisfactory Quality of service (QoS) to their customers is an important aspect for businesses these days. It is therefore essential that software systems can meet non-functional requirements in order to satisfy the business goals. Due to complexity of applications, highly varying request rates and resource scarcity, maintaining the desired QoS levels is not a straightforward task. The manual effort for system administration to maintain these systems to achieve business objectives is time consuming, error prone and costly [1, 2]. During last one and half decades there has been increasing interest in control theory to provide a solution to QoS performance management issues in complex software systems. Mainly application of control theory in web server systems[3, 4], cache and storage systems [5, 6], load balancing schemes [7], and data centers [8] can be highlighted as such efforts. However, all these approaches rely on single fixed or adaptive controller to provide the control objectives. We argue that the non-linear discontinuous nature of performance in complex computer systems necessitates an approach that can accommodate multiple controllers and selection methods. Appropriate controller would be selected depending on the configuration of the system, the performance of the components in the system and the contingencies environment.

In this work we propose to apply a control theoretic approach called Multi Model Switching and Tuning (MMST) adaptive control [9] to achieve QoS performance requirements and disturbance rejection objectives in software systems. At the same time we review the suitability of this approach under requirement changes, component and service failures and infrastructure (hardware profile) changes. These concerns are not addressed sufficiently by the approaches found in literature so far.

In Section 2, we introduce fixed and adaptive control schemes. Section 3 outlines MMST adaptive control scheme. In Section 4, we critically evaluate the applicability of MMST in QoS performance control using a model problem.

II. NON-ADAPTIVE AND ADAPTIVE SCHEMES

Figure 1. Block diagram of a Adaptive and Non-adaptive schemes

Figure 1 illustrates the Non-adaptive (rounded rectangle) and Adaptive control schemes. The objective of the controller is to maintain the measured output (controlled variables) of the system sufficiently close to the set point (reference value). To enable this, controller has to come up with the control input which will adjust a parameter in the software system. In the non-adaptive scheme, controller algorithms and parameters of the controller stays static over the entire period of operation. This makes it difficult to provide satisfactory control when large disturbances in the environment or the operating conditions of the system changes. To address that limitation adaptive control has this additional control loop that can change the controller algorithms and parameters according to the current environment and system operating condition. The Estimation component creates a mathematical model between measured output and the control input at runtime. This model is used by the Controller Design component to adjust the parameters of the controller at runtime to provide more accurate control according to the situation. However, performance of the adaptive control may degrade if there are sudden and large changes in the system or the environment,
leading to large control errors that could provide unsatisfactory control. For instance there could be time periods where most of the data retrieved from cache/disk or there is sudden high rate of requests. Due to these reasons model estimation component may come up with models with large estimation errors, which would ultimately affect the performance of the controller. Depending on environment and system conditions system model may change over time. In addition, request rates change continuously and unpredictably over time (e.g: high vs low request rates) which could map to different system models. These reasons illustrates that software system has multi model behavior, which requires mechanism to integrate multiple models and controllers in to software system for more accurate QoS performance management.

III. MULTI MODEL SWITCHING AND TUNING ADAPTIVE CONTROL

This section, we provide an overview of an approach called Multi Model Switching and Tuning (MMST) adaptive control [9] which support various multi-model model schemes.

![Figure 2. Block diagram of a MMST adaptive control](image)

MMST adaptive control was proposed by K. Narendra et al in early 1990’s to improved transient response of adaptive control systems [10]. When dynamics of the controlled system changes rapidly, adaptive controllers may not be able to track the parameters variations of the plant, results in unacceptable performance, also there are various estimation methods, and adaptive control techniques with various assumption, selection of the estimation method and controller design depending on the requirements is not a straightforward task [11]. MMST was proposed to address some of these limitations in adaptive control. It has been applied in some domains such as robotic manipulators and chemical process control systems [11]. Our work is focused on to apply MMST for QoS performance management of software system. MMST adaptive control is a concept inspired by biological systems [12]. Biological systems have the ability to select an appropriate action for a specific situation from a collection of behaviors, MMST uses the same concept by selecting the most suitable controller for the current environment, plant is in. The figure 2 shows the main components of MMST. Target system or the plant has control input $u$ and measured output $y$. There are $n$ number of models ($M_1, M_2, ..., M_n$) that may provide estimations for the system model, simultaneously. There will be $n$ estimates from these $n$ models denoted by $y'_{1}, y'_{2}, ..., y'_{n}$. Similarly, there are $n$ controllers, each corresponding to a model previously described. The responsibility of the switching algorithm is to select the appropriate model and corresponding controller pair based on some criteria that will improve the performance of the controlled system. There are multiple switching algorithms discussed in [12]. All of these algorithms are based on prediction error ($e = y - y'$). Switching rules for continues and discreet time systems are illustrated in [12] for this purpose. Two of the simple discrete switching algorithms are as follows. ($i = 1, 2,..., n$ and $k$ is the sample period)

$$J_i(k) = e_i^2(k) \quad J_f(k) = \sum_{r=1}^{k} e_i^2(r)$$

Given a time instance, model and controller pair with the minimum $J$ value is selected by the switching component to operate in the next time interval. So far we looked at the generalized methodology of MMST. Going further different types of multi-model schemes were evaluated and mathematical stability proofs were discussed in [13]. These multi model schemes are as follows.

**Type 1:** All adaptive models- in this scheme all the models (1, 2,..., n) are represented by runtime estimation models. The corresponding controller uses the parameter estimation to come up with the control input $u$. This scheme is computationally inefficient. In addition, if the environment remains unchanged for a long time, all the adaptive models will converge to same parameter neighborhood which reduces the advantage of having multiple models. In addition, when a sudden change occurs, models may not react to it rapidly without re-initialization of the parameters [13]. Moreover, initially adaptive models may not perform well, as it takes time for parameter values to converge.

**Type 2:** All Fixed models- in fixed models the above limitations are not seen. However, fixed models can only represent finite number of operating regions or environments. Apart from that, fixed models should capture the dynamics of different operating and environment conditions assuming that one of the models will be close to the plant model. To satisfy this assumption we may have to build a large amount of fixed models.

**Type 3:** One Adaptive model and one Fixed model – here initially fixed model will be selected most of the time since adaptive model takes time to converge at start. However, when adaptive model converges, it will outperform fixed model most of the time. This scheme is simple and addresses some of the limitations in above two schemes.

**Type 4:** Two Adaptive model and n-2 Fixed models-- in [12, 13], this scheme is regarded as one of the best schemes after the simulation results. From priory knowledge of the system operating conditions and environments, n-2 number of fixed models can be designed. Then one of the adaptive models will be running free of interference to capture the dynamics not captured by the fixed models. This model will improve
the accuracy of the system [13]. However, if a sudden disturbance occurs this model may not converge fast enough due to the delay of convergence for sudden disturbance, inherent in adaptive models. The other adaptive model is configured as a parameter re-initializable model to resolve this limitation somewhat. The main purpose of this model is if one of the fixed models were chosen by the switching scheme in successive time periods, this model could be regarded as the model that is close to the current system model. Using the parameters of that fixed model, the adaptive model parameter can be re-initialized so that systems can respond faster to sudden disturbances. Here fixed models have to be designed after analyzing the system and prior knowledge to achieve desired objectives. Interested readers are referred to [12, 13] for the details of stability proofs of these multi model schemes to guarantee that system will not be unstable due to switching and tuning behavior of them.

IV. MMST ADAPTIVE CONTROL FOR QOS MANAGEMENT

Znn.com is one of the model problems listed in Software Engineering for Adaptive and Self-Managing (SEAMS) exemplar website to compare different approaches proposed by the community [14]. Znn.com is news provider site with the objective of maintaining the repose time of the requests within a certain threshold when sudden large number of requests arrives (also called as Slashdot-effect) to avoid customer dissatisfaction. To make this possible system provides some reconfiguration capabilities (effectors) such as changing the number of servers in the pool or by switching the content modes (textual or multimedia). This model problem has different environmental conditions and operating regions. When the system is facing Slashdot-effect, the model of system may differ compared to facing low request rates. Using this prior knowledge we can create two fixed models and implement two controllers with different gain values. However, if we can build some more fixed models to capture the dynamics under different conditions, more effective MMST control can be implemented. The number of models largely depends on the performance overhead caused by them as well. The above models capture the dynamics of the environment conditions. There is several system operating regions as well. For instance when system is in textual content mode the model may differ from when it is in multimedia mode. In addition, there may be different models when server pool contains low number of servers and high number of server. From this discussion we can see the existence of multi-model behavior in this Znn.com model problem. MMST approach is applicable in this problem to achieve (response time) regulation objectives of Znn.com news provider. Availability of prior knowledge about environmental conditions and system will be really useful for the MMST integration in to Znn.com system architecture. Now, to summarize what we discussed, we can say that multiple models and controllers are required when there are dynamic and unpredictable disturbances, when there are multiple operating regions in system it-self, when there are possibilities for provisioning or configuration changes and software component failures. This raises lot of questions like how many models and controllers are needed, what are the models and controllers we can use, how can we switch between them, finally will it provide stability guarantees and performance improvements for the QoS management in software systems. MMST adaptive control is proposed to answer some of these questions. Firstly, it provides systematic way to integrate multiple models and controllers in to software systems and switch between them according to the situation. Secondly, if we have some priori knowledge about system operating regions and request rates we have the option to select one of the schemes discussed in section 3. The number and appropriate model estimation method for a specific application depends on the nature of the system, disturbance and prior knowledge available about the system. It has also well defined stability proofs.

In conclusion, the above discussion provides us some convincing reasons to apply MMST adaptive control in software systems. We are hoping to prove this concept using a case study as future work.

ACKNOWLEDGMENT

I would like to thank my supervisor Dr Alan Colman for his comments.

REFERENCES

Adaptive Traffic Assignment with Self-organizing Multi-Agent System
Syed Md. Galib
Swinburne University of Technology, Melbourne, Australia
Email: ggalib@swin.edu.au

Abstract – In this paper, I propose a self-organizing Multi-agent system for dynamic traffic assignment. The self-organization is done by Minority Game model which is mainly based on static cases in multi-agent systems. But my recent experimental results show that it is possible to have a model for dynamic traffic assignment with this Minority Game.

Keywords-Self-organization; Dynamic Traffic assignment; Minority Game;

INTRODUCTION
Traffic assignment is the selection of routes between an origin (O) and a destination (D) in a transportation network. A route is a chain of links between an origin-destination (OD) pair. Drivers, driving from an origin to a destination, make their decision about choosing a route when they come to an intersection of several routes (for example, Figure 1). Generally, they take decisions based on their past experience if they are driving a known OD pair, for example, driving from home to office. Some use the same route as it became more familiar to them than the others, some drivers sometimes avoid a route as they experienced congestion previously and some drivers just choose a route randomly. But one common thing among the drivers is they generally share limited information with each other.

Figure 1: Example of an OD pair with two routes

Now if the navigation systems in the car support the drivers to choose a route and for a same OD pair, the navigators suggest the same shortest routes to everyone, then they will all end up on the same route and thus having congestion and delays. Thus the cars are needed to be assigned on the routes in a more coordinated way which can be achieved using multi-agent system. In a multi-agent system, the agents act independently and share limited information which is almost similar to drivers taking decision about a route. Each driver can be supported to navigate to their destination by a software agent (e.g. a car navigation system). We propose to model the traffic assignment problem with a multi-agent system, where each car is modeled as an autonomous agent that learns to coordinate the road choices in order to optimize the overall traffic and reduce the congestion. Our aim is to develop adaptive and self-organizing strategies.

RELATED WORKS
Traffic Assignment problem has been addressed by many researchers in a number of ways. Drivers’ route choice behavior is studied in [1] and [2] where the drivers are provided with the information such as travel time of the last chosen route, last chosen route, payoff of the last period, cumulative payoff, the number of current period, and in some cases, travel time on the non-chosen-route as an additional feedback. There the authors showed that the fluctuation persists till the end and when the additional feedback about the travel time on the non-chosen route is present number of road changes is decreased. But in there, the drivers require the information provided from an authority which is not practical. In [3], the authors did not assume the network equilibrium as a priori but they tried to find out why the equilibrium had reached or not. They found out that equilibrium has been reached due to the delusion of the drivers when the drivers believe that the path they are taking is the best path and their travel time can’t be improved by changing routes. They named this as ‘deluded equilibrium’. The solution to this deluded equilibrium has been addressed in [4], where the authors suggested that if the drivers change routes suddenly or randomly, this deluded equilibrium can be dissolved and otherwise they develop the habit of choosing the same route and thus freezing their behaviors. Fuzzy inference technique has also been used for route choice for a two route scenario in [5] where a fuzzy network loading algorithm assigns traffic to each route but here the independence of the driver’s is not present. As the drivers predict the cost before start commuting, they do not know exactly what the cost will be. To deal with this uncertainty and imprecision, fuzzy route choice model has been proposed by Henn for traffic management in [6]. This model discussed on traffic assignment which does not allow driver’s independence. Multi-agent architectures have also been applied for traffic management both in centralized and decentralized manner. In [7], Hernandez, Ossowski and Serrano compared both centralized and decentralized multi-agent systems for traffic management using knowledge-based reasoning. They used multi-agent systems for controlling the traffic and thus the drivers have to depend on the signals to travel. Dynamic traffic assignment problem has been solved using mixed method of real-time simulation and off-line optimization in [8] where the cars are assigned to roads by agents. Multi-agent system is used there but the agents are used to calculate optimal routes in off-line using Frank-Wolf Algorithm and also some dependencies on the system for signaling are present. So, almost everybody formulates the road network as a graph and tried to solve using multi-agent system which does not give independence to the drivers. No research has been performed which models the cars or the navigation systems as agents.

METHODOLOGY
There are some solutions to the traffic control and assignment but those have central authority to control the traffic and assign the traffic based on the information of

Figure 1: Example of an OD pair with two routes

O

D
roads. For example, UTOPIA/SPOT [12] uses a hierarchical-decentralized control strategy, involving local controllers to communicate with signal controllers as well as a central computer and SCOOT [12] uses data from vehicle detectors and optimizes traffic signal settings to reduce the vehicle delays and stops. There are also expert systems with Advanced Traveler Information System (ATIS) [13]. But these systems do not allow the drivers to act independently and sometimes, some other expert systems require the cars to have communication with a central computer which makes them less practical. And also these kinds of systems are very expensive to set up. We want to solve this traffic assignment problem in a decentralized and self-organizing manner so that the solution becomes more practical.

Multi-agent system is very popular to solve problems in decentralized fashion. To have self-organization among the agents in a multi-agent system, researchers are applying some game theoretic models. And the Minority Game (MG) model, introduced by Challet and Zhang [11], can be applied in this kind of scenarios. In MG, finite and odd number of agents take independent decisions based on the previous collective decision of all other agents but do not share any other information with any of them. The agents have memory ‘m’ which is the last ‘m’ rounds’ winning side and they have predictors to predict the next winning side and choose that side. The predictors of each agent are given scores if they predict right or wrong. The number of agents in MG is static. But, in dynamic environment, MG has not been used earlier and the behavior of the agents is not known. My aim is to use the model of MG in dynamic environment and then to propose a model to solve the traffic assignment problem in dynamic environment based on MG.

Some assumptions are needed so that the traffic assignment problem can be solved by the minority game theory. The agents are the navigation systems in the cars. And there are two groups of people traveling – one group is the regular ones who travel regularly and the other group is the occasional ones who travel from time to time. The first group, who travels regularly, has the information about the roads that which road was less congested. And the later group, who travels occasionally, has less information. The first group usually reaches in some kind of equilibrium which might take long time but the others generally take decision randomly due to the lack of information. This group can cause congestion and if this group can be assigned on the roads optimally (or near optimally), the overall system optimization can be reached. This can be done in a self-organizing way so that the agents take their own decision to cooperate with each other without sharing any information except the past history. Finally, one more assumption is, all agents are using the same algorithm that they take their decision using some predictors with the help of the past history of congestion on the roads.

**SOME RECENT RESULTS**

Figure 2 shows some results of some experiments performed. In Minority Game the efficiency measure is Volatility which is the variance of total attendances of all agents in each round divided by the number of agents. But in dynamic case, as the number of agents is varying, the relative variance is taken to calculate the relative volatility. This is done by dividing the total attendances by the number of agents in each round and then taking the variance of those relative attendances. The lower the relative volatility is the better.

![Figure 2: Relative variance of the collective actions of each agent at each round in Minority Game in dynamic case](image-url)
In Figure 2, top left figure shows the relative variances for static cases for 101, 201, 301, 401 and 501 agents. The top right figure shows the relative variances for 301 agents in static case and relative variances for dynamic case where the number of agents varies from 101 to 501 and 201 to 401. The dotted line in the middle shows the relative variances when 301 agents choosing sides randomly. The two figures in the bottom show the relative variances for 401 and 501 agents in static case and with them the dynamic case where the number of agents varies between 101 to 701, 201 to 601, 301 to 501 and 101 to 901, 201 to 801, 301 to 701, 401 to 601 respectively. In last 3 figures, the relative volatilities of the static cases are the lowest than those of dynamic case. That is obvious as the number of agents is known and static. And the dynamic MG is not much worse. But the random choice agents’ relative volatility is much worse than both static and dynamic case which means a lot of fluctuations are present in random choice even in static case.

FUTURE WORKS

The memory size of the agents has an important effect on agents’ coordination. Still it is not clear with the number of agents with respect to time. And all the researches have been focused on the static case for Minority Game. But there is no research for dynamic number of agents. So, a model will be presented for Minority Game with dynamic number of agents. And finally, the dynamic traffic assignment problem will be solved using the dynamic minority game model. Validation of the model can be done by using it in some real world scenarios shown in Figure 3.

CONCLUSION

The dynamic traffic assignment problem has been addressed by the researchers but in all cases, the independence of the drivers has been ignored. I propose this model to solve the dynamic traffic assignment with self-organizing multi-agent system where the drivers or the cars or the navigation systems are modeled as the agents which will allow the drivers to assign themselves to the routes for a specific OD pair where several alternative routes are available. Eventually, they will learn with time to self-organize so that the roads are used in an optimized way and their travel time and delay reduce significantly.

ACKNOWLEDGEMENT

I thank my supervisors, Professor Ryszard Kowalczyk and Dr. Bao Quoc Vo, very much for guiding me from the beginning of my PhD. I also thank Tino Schlegel for helping to understand various things of Minority Game and other resource allocation techniques.

REFERENCES


Figure 3: Some real world scenarios where Origin is O and destination is D.
User Interface Model for Artifact-Centric Business Processes

Sira Yongchareon
Center for Complex Software Systems and Services
Swinburne University of Technology
Melbourne, Australia
syongchareon@swin.edu.au

Abstract—The artifact-centric business process model has emerged as a promising solution for modeling data-intensive business processes, as it provides a highly flexible solution to capture business process dynamics. However, unlike traditional process-centric models, it makes difficulty for users to perceive how a business process operates from its declarative manner of process modeling. Aiming to tackle this problem, we propose to derive the user-interfaces for artifact-centric process models, and thereby enable users to intuitively perceive and interact with the business processes. In this paper, a three-layered architecture is presented to address the relationship between business processes, user interfaces and user roles, with the help of the artifact system and UI flow model.

Keywords: Artifact-Centric Business Process, User Interfaces

I. INTRODUCTION

Over the past years the artifact-centric process modelling approach has emerged with particular focus on describing the data of business processes, known as “artifacts”, by characterizing business-relevant data objects, their lifecycles, and related services. It is believed that this approach improves efficiencies in business transformation as it provides a robust structure of business processes [1-3]. The artifact-centric process modelling fosters the automation of the business operations and supports the flexibility of the workflow enactment and evolution [4].

Unlike traditional activity-centric approaches, the artifact-centric approach does not provide an intuitive workflow diagram for users and modellers to see and comprehend the process structure. In regard to human computer interoperability, this could be a drawback of the artifact-centric approaches. Further, this feature influences the development/configuration of step-wise user interfaces for the enterprise systems which rely on artifact-centric business processes.

Having a closer look at artifacts and user-interfaces, we have observed that on one side a user-interface acts as a media for users to view/input business artifact data and invoke related functions to let the process proceed; on the other side the artifact data and business rules decide which UI should be brought to the user. Motivated from the above observations, we generate user interfaces for artifact-centric processes by deriving both behavioural aspect (control flow relations between UIs) and informational aspect (related data in each UI) of user-interfaces from artifact-centric processes themselves. Technically, we propose the architecture consisting of two models: Artifact-centric process model and User-Interface Flow (UIF) model. The UIF model describes the constitution of user interfaces and the sequence of these interfaces. The UIF can be derived from the underlying artifact-centric model. The derived UIF can be further customized for different roles of users, and therefore different roles may have different user-interface flows. In summary, our work makes the following contributions to the research in artifact-centric business processes:

- Analyze the relations between artifact-centric processes, UIs, and user roles
- Facilitate the UIs derivation for processes with UIF models and algorithms
- Support the customization of UIs for a particular role of users

II. USER-INTERFACE ARCHITECTURE FOR ARTIFACT-CENTRIC PROCESSES

User interfaces play an important role in human system interaction. As the media bridging systems and users, UIs show users particular data of certain artifacts and also enable users to input/edit the data of artifacts and invoke related functions. Driven by underlying business processes, a user may go through a sequence of UIs which help users fulfill a certain procedure of the business process. Here, we consider two aspects of UIs: behaviour and information. The behavioural aspect presents the flows between UIs, while the informational aspect presents the required data that users must complete for each UI as to proceed to its following UI. In the traditional control flow oriented business process approaches, the UI flows can be easily implemented according to the activity flows. However, the artifact-centric process approach does not own the activity flows.

The artifact-centric processes are defined and structured around artifacts. Three components are required to operate a process: artifacts, services, and business rules. Artifacts are business entities or objects involved in business processes.
Each artifact contains a set of attributes and a lifecycle described by the evolution of its states. Business rules are highly declarative and are used to associate services with artifacts in a Condition-Action-Role style. Each rule is defined in such a way that if the condition is true then the service is invoked as well as the associated artifacts (output) for such service are updated with the change on their states.

From the declarative manner in describing artifact-centric processes, we can see that data required to perform services is explicit while control flows between services are implicit. This brings in the challenge to develop a mechanism to discover the flows of UIs from such processes. In addition, UIs are defined for users and users in the organization have different roles. The authority to perform an activity in the process is restricted by particular roles of users. Different roles may have their own views of UIs. For this reason, we consider the role-based characteristic of UIs. Based on these requirements, we propose the UI architecture for artifact-centric processes. The architecture consists of three layers: Artifact System, UIF-Base model, and UIF-Role model as illustrated in Figure 1. The Artifact System layer comprises of aforementioned three components of business processes. It corresponds to the artifact-centric process model.

The UIF-Base model layer contains conceptual UIs and the dependencies among them. UIs are constructed based on behavior and information of business processes in the artifact system. Each individual UI in the model contains a set of states and attributes of related artifacts defined and used within it. The flow from one UI to another UI may be restricted by the behavior of correlated artifacts used in both UIs. The model is highly conceptual therefore it does not provide any physical layout or structure of the user-interfaces. This layer logically presents user-interfaces and their relations derived from the artifact system. The UIF-Role model layer represents the role-based user-interface model. This model can be seen as a user view of UIF-Base model defining which user-interfaces are enabled for a particular role. The roles are basically defined by the business rules in the artifact system. This model guides UI designers to design and customize physical user-interfaces.

The UIF-Base model and the UIF-Role model correspond to the UIF model. The arrows in Figure 1 show the direction of transformations or derivations from a lower layer to an upper layer.

### III. Artifact System

In contrast to ordinary business process model, which is activity-centric and focuses on control flows, artifact-centric is structured around data in a process. We define the artifact system to describe the artifact processes. We use the term system rather than process as it is conceivable in the ability of capturing multiple processes.

The main part of the model is a simplified version adapted from the concept of state machine which is used for describing the object’s behavior in a system [6]. The model consists of a fixed number of concurrent finite state machines of artifacts, called artifact machine, that have input events triggered by invoking services associated with the transitions of the artifact machines. One artifact has only one corresponding state machine. If the context is clear, we can say a machine to mean an artifact machine. The synchronization between machines is achieved by associating a guard with the transitions. Guard conditions are Boolean combinations of states of other machines in the system. By this means, transitions fired in one machine may affect state conditions of other machines. If guards in different machines share the invocation event of the same service, the transitions occur simultaneously.

### IV. User-Interface Flow Model Generation

In this section, we describe the terminology and constructs in User-Interface Flow Model (UIF model) and also propose an approach for generating the model from a given artifact-centric business process model. The UIF model comprises (1) the inputs of the interface required from users and (2) the flows or the relations among interfaces (UI). The model is at conceptual level; hence it will not represent the actual components and their structure in a user-interface, e.g., physical structures or layouts of the page, forms, inputs or buttons are not described.

Figure 2 shows the components and structure of the UIF model. The round-rectangle represents a User Interface Container (ε) or UIC. A single UIC may contain zero (for the final or initial UIC) or multiple interfaces (for normal UIC). The Interface (b) can be seen as a form comprising a set of input fields that user must fill in data to make the form completed. The interface also contains its corresponding artifacts and their current state if the interface is active. The Interface control flow (f) indicates that once the interface with all required data has been submitted, the action, e.g., service, corresponding to such interface is performed and the following UIC then becomes active. The UIF starts at the initial UIC and terminates when it reaches the final UIC.
To derive the UIF model from the artifact system, two main steps are required: (1) generating the interfaces and their control flow relations for constructing the behavior of the model and (2) mapping the required artifacts and their attributes for constructing the information of the interfaces.

V. RELATED WORK AND DISCUSSION

In the context of business process modelling, Küster Ryndina, & Gall [8] established a notion of business process model compliance with an object life cycle. They also proposed a technique for generating a compliant business process model from a set of given reference object life cycles in forms of state machines. Redding et al. [9] conducted a similar work, where they proposed the transformation from objects behavior model to process model by using the heuristic net for capturing the casual relations in the object model. Compared with our work, their transformations use an object behavior model as input, while our work uses the artifact process models. In addition, these approaches are different from ours in such a way that they do not consider state dependency but we do. Reuner and Schreil [10] proposed an approach to generate business process model by using object life cycles integration. The business process is also presented as an object lifecycle capturing the overall behaviour. In contrast to their approach, we integrate state machines of different objects to generate UI models.

In the area of user-centric aspect in business processes, both Sousa et al. [11] and Sukaviriya et al. [12] presented a model-driven approach to link and manage software requirements with business processes and UI models. With their approaches, a process model is mapped to a UI model, thus change propagation can be managed more efficiently. Guerrero et al. [13] applied the similar concept for developing UIs corresponding to a workflow model. All these approaches considered the traditional activity-centric process model and proposed approaches to define the internal components and functionalities of the UIs at different levels, e.g., task-base model, abstract UI, and concrete UI. In comparison with these approaches, we considered the artifact-centric model in this paper. The generation of UIs from artifact-centric business processes is much more challenging. We proposed the architecture at a higher level that provides a highly-cohesive bridge between the operational back-end system of business processes and the front-end UI system and it can be further extended to cover those detailed levels. Furthermore, our architecture provides a feature to support role-based customization and configuration of UIs.

VI. CONCLUSION AND FUTURE WORK

This paper has proposed the user-interface architecture for artifact-centric business processes. The UIF-Base and UIF-Role models are defined as to support the user-interface generation for enabling the visualization of the process and information flow to process users. Apart from that, the UIF models intuitively represent what information is required during the process and how user-interface designers can use this generated conceptual model to build concrete user-interfaces. In the future, we plan to improve the model for supporting more specific user-interface requirements, e.g., attribute settings, interface views, and interface for tracking artifacts.

REFERENCES

On-the-Fly Coordination of Automated Negotiations

Khalid Mansour
Faculty of Information & Communication Technologies
Swinburne University of Technology
Melbourne, Australia
mwmansour@swin.edu.au

Abstract—The coordination process is an effective mechanism in managing actions of different parties working together for a common goal. This paper addresses the importance of coordination during concurrent multi-bilateral negotiations. When a group of agents are negotiating for a certain purpose, their decisions need to be aligned in a coherent manner since the totally independent actions of agents may not be successful in fulfilling their goals. This paper focuses on the coordination problem description and presents some simple experimental results. The results show that applying coordination mechanisms can significantly improve the negotiation outcomes.

Index Terms—negotiation; coordination; interdependencies

I. INTRODUCTION

The need for a coordination mechanism arises when multiple agents (or actors) are working together to achieve a common goal. Accordingly, it is intuitive that the independent decisions of agents produce unpredicted results. This research investigates the problem of coordination in multi-agent systems. In particular, we focus on coordinating the decisions of agents during negotiation.

We distinguish between two types of negotiations that require coordination. The first one is called multilateral negotiation. In multilateral negotiation, many parties are engaged in negotiation at the same time for the purpose of achieving an agreement for all parties. For example, the negotiation between different nations for the purpose of reaching a political agreement is called multilateral negotiation.

The second type is called multi-bilateral negotiation which means that there are multiple instances of negotiations and each instance involves only two parties. When multiple instances of negotiations are working together for a common purpose, their decisions need to be aligned in a coherent manner since the totally independent actions of the instances may not be successful in fulfilling the planned goal. For example, if agents $A$ and $B$ are working together for a common goal, then an agent $A$ might not be able to accept an agreement before an agent $B$ does. This research investigates the second type (i.e., multi-bilateral negotiation).

The negotiation process describes a state of interaction between two parties (e.g., a buyer and a seller) over one or more issues. Correspondingly, this research presents the coordination from a buyer’s point of view over one or more issues. This research focuses on the problem of managing concurrent negotiations in an environment characterized by limited critical information, where the buyer agent seeks to maximize its collective gain out of negotiations.

II. RELATED WORK

Negotiation is an effective decision making mechanism and dispute neutralization method that can be used by opponents having conflicting interests [10]. The basic form of negotiations is the bilateral (one-to-one) form where one agent negotiates with other agents over certain issue(s). The one-to-many form of negotiation is another form of agents interaction. In one-to-many negotiations, one agent negotiates with many other agents concurrently. The one-to-many form of negotiations is more complex than the one-to-one from because in the first form an agent needs to interact with multiple agents concurrently and that implies a need for coordinating the various responses of that agent.

Some studies investigated the one-to-many form of negotiation, for example, [6] [9] describe a scenario where the buyer agent is interested in negotiating with many sellers over a single type of good or service. There are more complicated scenarios where the aim is to procure a set of different services or goods from different providers. [1] Describe an architecture for procuring a set of different services and elaborate on the functions of the coordinator in such situation, but the study does not propose certain methods or mechanisms for conducting the coordination process.

[3] and [4] describe two different frameworks of one-to-many negotiation form. [4] illustrate a special scenario in e-commerce, but the proposed framework does not describe a real negotiation scenario between agents in terms of offers and counter offers generation, and the process can be described as a method of searching for the best matching offer amongst the sellers agents. While [3] describes a more general situation of real negotiation scenario for the service level agreement (SLA) negotiation. But when a negotiator agent is capable of negotiating with many other agents, it means that the scenario describes one-to-many negotiation, which implies that the framework has either two levels of coordination or each negotiator agent should negotiate with one service provider at a time. The framework was unclear about this point.

The behavior modeling of an opponent [8] depends on two sources of information: before the start of negotiation and during negotiation. Firstly, The available information before
In the automated negotiations context, we define the coordination as the process of synchronizing the decisions of agents or threads during multi-bilateral concurrent negotiations in a way to maximize the collective outcome gain of the agents’ decisions or to minimize their total loss.

In our case, we model the coordination process as a system which has inputs and outputs. For example, surplus in resources, reservation values and deadlines are examples of the inputs. The outputs of the coordination process determine the effectiveness of the process in achieving its goals and the efficiency of the process in utilizing its resources. Various metrics can be used to determine how effective and efficient the process is. For example, if the goal of negotiation is to achieve the highest possible number of agreements with opponents, then the effectiveness of the agreement can be measured by the agreement rate and the efficiency can be measured by the utility attained.

The difference between the reservation value and the agreed upon value is called a resource surplus. The coordinator decides how to use the surplus(s) in resource(s). The reservation values and deadlines are the negotiation constraints.

The following is a formal description to the different variables that can affect the coordination process:

For the purpose of clarity, we assume that a buyer agent (b) is negotiating with a set of sellers (S). The following model presents a buyer’s perspective. The other way around can also be build in a similar way. We assume the following:

1) The buyer consists of a coordination entity (co) and a set of negotiation threads (D),

   \[ b = \langle co, D \rangle \]

   \[ D = \{d_1, d_2, \ldots, d_m\} \]

2) The number of threads equals the number of sellers (|D| = |S|) and each thread corresponds to exactly one seller (D \( \xrightarrow{1-1} \) S).

3) Since the negotiation issues are the subject of negotiation, let I stands for the negotiation issue set.

   \[ I = \{i_j\}_{j=1}^{n} \]

4) Let \( \alpha_j \in 2^I \backslash \emptyset \), where \( j = \{1, 2, \ldots, |2^I| - 1\} \), and each \( \alpha_j \) may represent a certain good or service (item). Each set of issues \( \alpha_j \) may represent the same or different items, for example, a book and a pen may have the same issue set of negotiation (e.g., price and quality).

5) Let \( d_j \in D \), \( d_j = \langle \alpha_j, t_{max}^{d_j}, w_j, g_j \rangle \) where \( j = \{1, 2, \ldots, |D|\} \). \( t_{max}^{d_j} \) stands for the deadline of thread \( d_j \), \( w_j \) stands for the weight of thread \( d_j \), \( g_j \) stands for...
the negotiation tactic or strategy of thread \( d_j \). The \( w_j \) actually reflects the importance of a certain item that the \( d_j \) is responsible for. Similar settings are applied to the set \( S \).

6) Let \( bd \) stands for a buyer threads subset such that \( bd \subseteq 2^D \setminus \emptyset \). The coordinator selects a subset \( bd \) such that \( \forall l_j \in bd, \bigcap_{j=1}^{\|bd\|} l_j = \emptyset \) & \( \bigcup_{j=1}^{\|bd\|} l_j = D \).

7) Each \( l_j \in bd \) corresponds to a distinct item represented by a set of issues \( \{ l_j \to \alpha_j \} \).

8) Let \( m_j \) stands for the number of service providers of a certain item, then \( |l_j| = m_j \).

IV. METHODOLOGY

We consider the coordination of direct multi-bilateral negotiations under incomplete and uncertain information. This research aims to build various coordination models for the automated multi-bilateral concurrent negotiations, since one coordination model might not be suitable for all negotiations settings and objectives. The design of each coordination model is scenario-driven. For example, the scenario of procuring a single good is different from procuring multiple distinct goods while negotiating with multiple providers. The different scenarios can be seen from two different perspectives, a buyer’s perspective and a seller’s perspective. We need to investigate the effect of the independent variables such as the available information about the opponents on the dependent variables such as the utility of negotiations when using different coordination mechanisms (see figure 2).

The empirical experiments are the main validation method for the proposed models. Different metrics can be used to evaluate the proposed methods and techniques such as utility, agreement rate, Pareto optimality, etc.

V. EXPERIMENTAL RESULTS

![Percentage of Agreements](image)

Fig. 2. Agreement rates with two different surplus distribution methods at two different times

Figure 2 compares between two distribution methods. D1 & D2 stands for the starting of surplus distribution using the on-demand method at the beginning of negotiation and at the middle of negotiation respectively. W1 & W2 stands for the starting of surplus distribution based on the weight at the beginning of negotiation and at the middle of negotiation respectively. The initial experiments show that using W1 & W2 deliver better results than D1 & D2. The results show that the time of starting surplus distribution has no significant effect on the overall results. We need to run more experiments to verify the initial results.

VI. CONCLUSION

This paper describes the coordinated negotiation research problem and show some simple experimental results. Coordinating the actions of agents is inevitable activity during negotiation and should be managed in a coherent manner. The future work will focus on using learning and prediction methods to model the behavior of the opponents and the environment of the negotiation to predict the actions of each individual opponent which affect the coordination decisions in terms of strategy selection and redefining utilities of agents. We will start with a single issue, then the focus will be on multi-issue negotiation and coordination. Whether the issue of negotiation is continuous or discrete can affect the coordination approach.

The future work will investigate the coordination techniques with respect to various inputs and scenarios.

ACKNOWLEDGMENT

Many thanks to my supervisors, professor Ryszard Kowalczyk and Dr. Bao Yo for their continuous help and support.

REFERENCES


Large-Scale Emulation of Enterprise Systems

Cameron Hine
Faculty of Information & Communication Technologies
Swinburne University of Technology
P.O. Box 218, Hawthorn, VIC 3122, AUSTRALIA
chine@swin.edu.au

Abstract—Enterprise software environments are heterogeneous dynamic systems which consist of many thousands of distributed nodes all interacting with one another. Developing software for deployment into such environments remains a significant technical challenge, in particular, it is difficult to test software in an environment that is representative of such variety, scale and complexity. In this article we outline an approach to representing large-scale interactive representations of such environments, for testing purposes. The popular current approaches are discussed and the advantages of our adopted emulation approach are argued. Progress towards realisation of this emulation environment is also discussed.

Keywords-large-scale emulation; enterprise system; protocol modelling; protocol conformance

I. MOTIVATION

Enterprise software environments are dynamic systems that typically consist of many thousands of distributed nodes interacting with one another via shared communication channels. In addition to the properties of distribution and scale, enterprise software environments are heterogeneous: individual systems are of different types, serving different purposes and communicating with one another using different protocols.

The compounding factors of distribution, scale and heterogeneity make developing high quality software for deployment in enterprise environments, a significant technical challenge. In particular, it is difficult to test enterprise software in an environment that is representative of the variety, scale and general complexity which can be expected in a production deployment. This kind of system testing is crucial to ensure that an enterprise software implementation can handle the scale and heterogeneity of environments it is likely to encounter when deployed onto various client infrastructures.

The remainder of this article will outline an approach to representing enterprise software environments for testing purposes which we dub emulation. We consider our approach to be emulation rather than simulation as our environment is open, rather than closed; the emulation will communicate with a real enterprise system under test, not a model thereof. In addition to representing an interactive variety of enterprise software environments, the emulation environment will be able to detect protocol conformance errors at runtime in enterprise systems under test. Protocol conformance testing at this scale has not previously been possible in enterprise environments and represents a significant contribution to the understanding of the interaction behaviour of enterprise software at scale.

The rest of this article is structured as follows: Section II outlines current approaches to representing enterprise software environments for testing purposes as well as work on modelling communication protocols. Section III details our emulation approach and discusses the current status of the work. Section IV describes how we intend to validate the work and summarises the expected research contributions to come out of the project.

II. CURRENT APPROACHES AND RELATED WORK

There are numerous existing tools and approaches to providing an interactive representation of distributed systems for testing purposes. These approaches include, but are not limited to, performance testing tools, network emulators, virtual machines, and ad-hoc programmatic techniques. Each of these approaches and tools have a certain concrete purpose for which the resulting environment representation is sufficient. Unfortunately, no existing approach fully addresses the needs of an interactive enterprise environment for testing purposes, as will be illustrated by what remains of this section.

Performance testing tools, for example, help identify performance issues in a system under test by placing stressful load on that system. Tools such as SLAMD Distributed Load Generation Engine and HP’s LoadRunner are capable of representing many thousands of concurrent clients issuing requests to a system under test. Some performance testing systems are capable of providing sophisticated diagnosis of performance issues, this subsequently aids developers to identify and address these performance issues. Unfortunately, performance testing tools are geared towards representing active (client) systems and are less capable at representing more complex interactions exhibited by reactive (server) systems. The enterprise software systems this research is concerned with will often act as the reactive interaction partner. This means the environment representation will need to exhibit server patterns of behaviour, such as that of an LDAP [1] server for example.
Virtual machines such as those provided by VMWare Workstation [2] and VirtualBox [3] model and execute complete physical machines. These virtual machines are commonly used as the basis for testing environments as they provide useful administrative features such as machine state snapshot, which can be restored at a later date. Virtual machines scale to a certain extent, up to approximately twelve virtual machines can be run simultaneously on a single (high-end) physical server. Unfortunately, the models used by virtual machines are what we term heavy-weight, they capture a resolution of detail which is not absolutely necessary for enabling a variety of testing activities. The emulation approach we take is similar to the virtual machine approach, however the models we use are intended to be lighter-weight and capture only the details necessary for enabling testing scenarios. The assumption being that lighter-weight models require fewer computational resources to execute, thus enabling a larger number of models to be executed on a single physical machine.

Mock objects [4], [5] and other programmatic approaches allow developers to write small implementations of a systems behaviour such that it is sufficient for limited testing purposes. These approaches are typically hooked into the code of the system under test, on the other hand, the emulation approach we take allows the system under test to be treated as an opaque-box, requiring no access to the code of the system under test to be of use in a variety of testing activities. The benefits of this is twofold, (i) the enterprise system under test is exactly as it is in a deployment, and (ii) components for which the source is unavailable can also be tested.

Runtime Protocol Conformance: In addition to providing an interactive representation of an enterprise environment, the emulation environment ensures that certain faults in a system under tests protocol implementation will be detected. To this end, a model of enterprise protocols is required.

There are numerous protocol models each suited to different contexts. These models include formalisms such as finite state machines [6], type based approaches such as session types [7], [8] and regular types for active objects [9]. Unfortunately none of the existing formalisms were found to fully meet the unique protocol modeling needs of enterprise system protocols. This has lead us to creating a novel protocol model described in a paper entitled “Modelling Enterprise System Protocols and Trace Conformance” to be published in the proceedings of ASWEC in 2010. The model described within that work was shown to enable concise description of complex hierarchical interaction patterns common to enterprise system protocols.

III. System Modelling and Emulation

Enterprise system modelling and emulation is the approach we adopt to representing large scale enterprise environments. The key aspects of this approach are represented pictorially in Figure 1. The two central concepts in this approach are: (i) Modelling the interaction behaviour of real enterprise systems which the system under test would usually interact with, (ii) An emulation runtime environment which provides the infrastructure to execute these models and communicate with the system under test in place of the real enterprise environment.

The fundamental idea behind our approach is that the models of enterprise systems are light-weight; the computational resources required to execute a model of an enterprise system shall be significantly lower than those required to execute a real enterprise system. This lightweight characteristic allows the runtime environment to simultaneously emulate a large number of instances of an enterprise system model, appearing to the system under test, to be a large scale enterprise environment, despite the fact that there are only one or two physical machines doing the work. Importantly, the models are capable of describing both reactive and autonomous system behaviour, allowing the modeling of client as well as server type systems.

A key feature of our emulation environment is its ability to detect protocol conformance errors originating from the system under test. This makes it possible to check the correctness of an enterprise systems protocol implementation when it is scaled to interact with large scale environments of varying configurations.

This approach addresses the most significant limitations of the current approaches discussed in Section II. The models of enterprise systems allow for description of both reactive and autonomous behaviour, thus improving on the purely autonomous load generation approaches. The system under test is treated as an opaque box, this is an improvement over mocking approaches which require hooks into the actual code. Finally, the light-weight models provide a marked improvement in scalability over more heavy-weight virtual machine models. In fact the emulation approach could be deployed onto a number of virtual machines further improving the scalability.

It should be noted that the approach outlined in this section is not entirely theoretical. We have already demonstrated the viability of the approach in previous work [10] where finite state machines in combination with an immutable datapool was used to emulate up to 10,000 LDAP directories simultaneously. Moreover, work on modeling protocols and checking an implementations conformance to those models is to appear in the proceedings of ASWEC 2010. This work provides a basis for the runtime protocol conformance checking that will be present in the final prototype environment.
IV. VALIDATION AND CONTRIBUTIONS

There will be two forms of validation for the results of this project. Firstly, a prototype emulation environment will be constructed which will serve of a proof of concept; illustrating that indeed enterprise systems can be modelled in such a way that they can be emulated on a large-scale by a single physical machine. Moreover that the emulation can largely fool a system under test, at least so far as to enable useful testing to be conducted.

Secondly, a number of case studies will be performed where the prototype emulation environment will be connected to industry supplied enterprise software. These case studies will demonstrate that the emulation approach can indeed be used to execute enterprise software in environments of significant scale and complexity. Additionally, detection of protocol conformance errors in industry software will demonstrate the power of the approach in detecting faults in real software which are difficult if not impossible to replicate using existing techniques.

This work will enable both industry and academia to conduct large-scale software system testing of a kind not possible through current techniques. Significant contributions have already been made regarding the understanding of modelling application-layer protocols and verifying an implementation’s conformance with respect to those protocol models.

ACKNOWLEDGMENTS

Many thanks to my supervisory team: Jean-Guy Schneider, Jun Han, Lars Grunske and Steve Versteeg, for numerous invaluable consultations. Also many thanks to CA for their ongoing financial support of the work.

REFERENCES


The Resolution of Symbol Meaning in Multi-agent Systems
Autonomous Emergence and Alignment of Meaning in Multi-agent Systems

Wojciech Lorkiewicz
Wroclaw University of Technology / Swinburne University of Technology
Wroclaw, Poland / Melbourne, Australia
E-mail: wworkiewicz@swin.edu.au

Abstract— Autonomy of a multi-agent system, in relation to the external environment, can be greatly extended through the incorporation of a language emergence mechanism. As such, allowing the population of agents to autonomously learn, adapt and optimize their semantics; and introducing a flexible and effective communication language shared among the entire system. However such a mechanisms must guarantee that an individual is able to correctly identify the referent of a newly identified symbol; must guarantee that the population develops a consistent and common substances of symbols, etc.. Due to these problems the design of such a mechanism is a complicated and complex task. Therefore, currently only the language game mechanism has been broadly studied, and despite several limitations successfully applied. In this research we try to address these limitations in three basic cases of semiosis, each representing a possible scenario of language emergence in a multi-agent system.

Keywords-component; language grounding; communication emergence; cognitive agent; multi-agent system

I. INTRODUCTION

Over the past few decades computational models of language have focused only on symbolic explanation of linguistic meaning. Artificial systems incorporated fixed symbol-meaning assignments, highly limiting the autonomy of an individual and causing a brittleness of the system in the case of unexpected or unforeseen conditions. These systems assumed a set of arbitrary "physical tokens" manipulated on the basis of purely syntactic "explicit rules" [5]. Resultant, the meaning of symbols could be only defined in terms of other symbols and as such producing meaningless circular definitions. Whilst, humans are less hindered by such circularity as they ground words in their physical experience of the external world [1]. This raises the question whether the knowledge of such a system should be pre-programmed by an external designer, or rather should the concepts be made intrinsic to the system?

Harnad [5] proposed that symbolic representations should be grounded bottom-up, creating a coupling between environmental experience and the symbol. As such the grounding can be formalized as defining a method that constrain, over sensorimotor data, the use of a given symbol, e.g. a classifier, a perceptual/pattern recognition, etc. If effective method is available, then the symbol can be considered as grounded. Following Brooks [2] “to build a system that is intelligent it is necessary to have its representations grounded in the physical world”.

Obviously a grounded system should have numerous advantages. First of all such system should be robust, as its performance should not degrade in case of unexpected and unforeseen situations. Secondly it should be adaptive, as it would allow unattended changes in the language itself. Lastly it should be more effective, as its performance should improve with time. Furthermore, the ability to understand the meaning of a symbol should be among the mental states of the system, i.e. the symbols must be grounded in the external world that they refer to.

In order to allow agents to communicate with each other the individually emerged and grounded representations should be further aligned and shared by the entire population. As such the population of agents should autonomously learn, adapt and optimize their semantics. The superior goal of the resolution of language symbols meaning in a multi-agent systems is the resultant formulation of consistent and common substance of symbols, i.e. conventionalized symbols.

II. EXISTING APPROACHES AND THEIR LIMITATIONS

Recently several approaches to grounding problem have been proposed, see [4][7][8] for overview. Cangelosi and Parisi described a model in which an innate communication system evolves in a population of neural networks that forage for mushrooms, i.e. signalling poisonous/edible mushroom. However NN approach is faced with the problem of providing the learning set, as Christiansen and Chater (1992) argues it will always be parastatic to the system. An alternative approach was introduced by Marocco et al. (2003), where a genetic algorithm was used to evolve a lexicon to coordinate interactions of a robot arm with two different objects: a sphere and a cube. Their approach focused on the vertical evolution of the language, i.e. the evolution between populations, whilst neglected the inner population dynamic – horizontal evolution. However, most interesting research was carried in a series of successful experiments by Vogt and Steels (since 2000), where robots developed a lexicon using the language game model. Thus providing support for considering the language game as a complex adaptive dynamical system.\(^1\)

In the language game model (LGM) a population of agents tries to develop a shared lexicon, a set of associations between words (strings of characters) and meanings (features of objects), using communicative acts. In short, LGM is a routinized interaction between two agents, where one of

\(^1\)The references are omitted due to limited space.
them acts as a speaker, labelling an object in its attention view, whilst the other acts as a hearer, trying to identify the object in the context. Vogt and Coumans in [6] described three types of language games, using shared attention - observational game, using corrective feedback - guessing game, and using no feedback nor joint attention - selfish game (cross-situational learning).

Despite experimental success the LGM has still several severe limitations. First of all the game is always in pairs, i.e. any interactions between multiple agents are impossible. Further, it is limited to only three possible situations, i.e. case with positive feedback, full feedback and with no feedback at all, and requires an additional form of pre-assumed non-verbal communication, i.e. pointing and acknowledging. Thirdly, the interactions between agents are limited to a single one-step procedure, i.e. allowing only basic interaction between individuals. Additionally all of the current LGM approaches used a representation enforcing the one-to-one mapping between the sign and the meaning, i.e. introducing synonymy and requiring additional synonymy dumping mechanisms. Moreover, as shown by De Beul and Bergen (2006), “when there is competition between general and specific meaning – the general is dominant” (Dominance of Generality). Resultantly all the emergent languages have a tendency towards a holistic language and as such disallow any compositional structures to emerge. Further all observations made by the agents are assumed to be complete and flawless, i.e. there is no uncertainty in the agent’s perception. Last but not least the LGM is a static and innate mechanism of language emergence. ²

III. PROPOSED APPROACH

Assumed agent is equipped with a given set of sensors that allow it to register certain signals from the environment, i.e. attributes currently exposed by the objects. From the phenomenological point of view the attributes can be perceived as different cognitive modalities that the agent is able to experience, whereas set of all attribute values define the possible space of values for each modality, see [11] for details.

![Figure 1. A) Individual Semiosis; B) Population Semiosis; C) Cross-Population Semiosis.](image)

Developing the complete mechanisms of language symbol alignment requires dealing with three basic problems, i.e. **individual language emergence** (Fig.1 A), where a single agent is aligning with the general population semantics, **population language emergence** (Fig.1 B), where each single agent is forming the general population semantics, and **cross-population language emergence** (Fig.1 C), where two or more populations align their own semantics.

A. Individual Semiosis

An individual is shaping and aligning its language with a distinguished external source of meaning. A single agent $A_0$ is interacting with a group of mature agents $A$ (with a predefined meaning of language symbols) in order to learn how to correlate the language symbols. These symbols are imposed by the mature population and can be correlated by $A_0$ with empirically perceived external states of the environment that are the source of grounded meaning. As such through numerous integrations with the mature population the single agent is able to identify the source of meaning of for each language symbol.

B. Population Semiosis

More generally the problem of shaping the meaning of language symbols can be deliberated without the need of a distinguished and a predefined population of mature agents. The process is then distributed among individual agents, where each is autonomously developing and adapting its own personal semantics. As such each agent is individually grounding the meaning of symbols, and aligns it with the entire population. The former allows agent to autonomously perceive the environment, whilst the latter allows the population to reach a common understating.

C. Cross-Population Semiosis

The third case is when two mature populations, with differently predefined semantics (or differentially developed as of the Group Semiosis), are set together in a common environment. Treated as a whole the collated populations may have various meanings assigned to the incorporated language symbols, and in order to communicate their semantics should be aligned, or may have different language symbols, and both dictionaries should be related to each other. The novelty of this approach lies in the fact that all of the aligned ontologies share the same environment, which is common to all populations. This common background of reference can lead to a proper correlation between ontologies, as the existence and state of the external world is objective rather than subjective.

IV. LANGUAGE GAME MODEL EXTENSIONS

In order to overcome the limitations identified in section II, two fundamental extensions to LGM are introduced. First is the study in the area of meaning representation, focusing to allow more flexible and rich form of representation, and the second on enriching the Language Game Model, to a more advanced, flexible and exact mechanism of language emergence coordination.

Meanings should be developed and not predefined, i.e. grounded in agents sensorimotor experience and inner population interaction. Further they should be flexible, i.e. not representing the one-to-one mapping, and dynamic, i.e. able to evolve due to agents interactions. Proposed meanings

---
² For additional limitations of current approaches please refer to [8], [9] and [10].
should also be internal to the agent, i.e. private and not imposed by a centralised mechanism of coordination, and cannot be directly communicated.

Enrich Language Game Model (LGM) to be able to cope with proposed flexible meaning, i.e. incorporate assumed flexible meaning representation to already existing language game mechanism. Original formulation of LGM is only between multiple individuals and should be extended to introduce games between more than two agents, i.e. multiple speakers, multiple hearers, mixed multiple speakers and hearers. Additionally LGM is introduced as a single stage interaction, therefore it is limited to a very simple communication, as such a multistage protocol could drastically lower the uncertainty. Enrich the single stage protocol interaction in order to gain more precise knowledge. Additional research should be performed on the possibility to incorporate different LGM protocols, i.e study the other possible scenarios.

V. CURRENT RESEARCH

Currently the simplest case of individual semiosis has been implemented in the JADE framework and basic consistency study has been performed. The environment consisted of a given set of objects. Each object in a given time point exposed its current set of properties, see [11] for details. The whole system was dynamic, as the objects properties evolved during time. In the experiment a new agent was introduced to a population with already predefined language symbols semantics, where it was able to perceive the objects and communicate with the population. Based on two algorithms for acquiring language semantics: CSSL (without biases) and Modified CSSL (with mutual exclusivity bias), the agent was able to acquire the mapping between symbols and meanings that were used in a population.

Other undergone research focused on the possibility to represent other agents’ meaning, as an approach to develop representations for embodied ontologies used by other agents. Here agents were assumed to possess their own systems of meaning captured by their private ontologies. However, in order to communicate successfully with other agents they have to know what meaning could be assigned by the other agent. By analysing correlations between incoming language symbols and states of external world an agent were able to develop internal reflections of meaning assigned to the used symbols by the senders. As such the agent was able to create, for every other agent in the population, its personal reflection of embodied ontology. In the performed experiments the agents used this knowledge to divide population to consistent language communities in order to improve the overall communication quality, see [11] for details.

VI. FUTURE RESEARCH AND SUMMARY

Despite of early critique the autonomous language emergence is currently gaining additional interest in the research community, as it offers a plethora of open questions and research opportunities. As briefly sketched, the problem of aligning the meaning of language symbols should not be treated only at the theoretical level of deliberation, where a population of agents dynamically adapts the communication language to suit the shape of external world, assumed perception mechanism and intra-population interactions. As this problem is fundamental for several application areas, e.g. robotics (embodied agents) where each agent perceives the environment and communicates with other agents, human-computer interfaces, smart sensor networks or information sharing environments (e.g. social portals, blogs, wikis, etc) where objects are correlated with a set of labels, keywords or tags. The superior goal of the resolution of language symbols meaning in a multi-agent systems, i.e. autonomous emergence and alignment of meaning, is the resultant formulation of consistent and common substance of symbols, i.e. conventionalized symbols.

Future work will focus mainly on the development of effective algorithms that could realise all three cases of semiosis. Secondly, the language game model will be enriched to a case of multiple stage and multiple agent interaction. Thirdly thorough investigations concerning the necesscity of additional assumptions will be performed. Finally, it is assumed that the whole system will be implemented and experimentally verified.

REFERENCES


Abstract—Wireless sensor network is a fascinating technology that promises to reshape the existing communication networks. The potential real applications are vast and the research challenges are multiple. To optimize the usage of energy resources and share the wireless medium efficiently stand as the most important research issues. In this extended abstract we present a systematic literature review on the topic, identify the research problem and explain the motivation. In addition, the intended methodology and plan for validating the work is described.

Index Terms—Wireless Sensor Networks; Analytical Modeling; CSMA-CA Optimization; IEEE 802.15.4

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are an emerging technology for low-cost, unattended monitoring of a wide range of environments [1], made possible by recent advances in microprocessors and low power radio transceivers. WSNs can be used for a broad range of applications areas, such as national security, surveillance, military, health care, environmental and building monitoring, industrial automation, etc.

Sensor nodes deployed in a field can automatically organize themselves to form ad hoc multihop wireless network. The communication is carried out between all nodes and with a central coordinator, usually named sink. Several surveys [1]-[6] discuss various aspects on wireless sensor networks and give a comprehensive review of many developments and challenging issues that are needed to overcome and present the solutions proposed up to now.

The main characteristics of WSNs were outlined in [2]. The number of sensors in a network can be of several orders, hundred or thousand, are densely deployed and are prone to failures. Hence the network topology changes very frequently. The communication is mainly of broadcast nature. Sensor are limited in power, processing capabilities and memory. These characteristics determine a series of general design factors. The tolerance to failures, scalability of the network and use of wireless medium for transmission are some of the factors. However, the management of power consumption is the most important element to consider due to the energy constraints of a battery-powered device. Researchers are approaching these challenges with the introduction of enhancements in every layer of the OSI model and also constructing parallel planes for issues such as task, mobility and power management [2].

The primary objective in WSNs design is to maximize network lifetime. A sensor node is generally composed of a sensing unit, processor and a radio transceiver. The study [7] found that the radio transceiver is the most energy depleting unit in a sensor node and breakdowned the power consumption of a widely use commercial transceiver (CC2420) in each operation state: shutdown, idle, transmit and receive. We can realize that the medium access control (MAC) sublayer plays a fundamental role in the energy management since it controls the transceiver operation. The MAC-specific sources of energy waste are overhearing, collisions, overhead and idle listening. To solve these issues several new MAC protocols has been designed for WSNs.

The release of IEEE 802.15.4 specification[8] standardized a MAC protocol for low data rate wireless sensor networks which is based on a combination of random and scheduled access. The specification incorporated power saving mechanisms in the two modalities of operation: beaconless and beacon-enabled mode. In the case of beaconless every node accesses the medium using an unslotted Carrier Sense Multiple Access with collision avoidance (CSMA-CA) scheme. For beacon-enabled the coordinator establishes a duty cycle period that allows sensors to enter in shutdown state, and hence save energy. The active part of the cycle is divided in a contention free period (CFP) and a contention access period (CAP). In CFP node follows a scheduled access with the use of guaranteed communication time slots and in CAP a CSMA-CA scheme is employed.

There exist other protocols [10][11][12] for medium access control. Each serves specific application scenarios and is limited in terms of modes of communication. Some of those protocols provided the foundation for the IEEE 802.15.4 standard. For these reasons, academic and industry have embraced the IEEE 802.15.4 specification, which can be used in many environments.

II. PROBLEM DEFINITION

Studies on real WSN deployments [19][20] insist on the importance of finding mechanisms to reduce the energy consumption of sensor nodes. This objective can be achived by means of analyzing and enhancing the IEEE 802.15.4 MAC protocol. With these considerations in mind, the first aim of this project is to characterize the performance of the IEEE 802.15.4 MAC protocol operating in a beacon-enabled mode and under a carrier sense multiple access with collision avoidance (CSMA-CA). To do so, we are building an analytical model that captures the behavior of a single sensor.
node in a general star topology network. The model will allow to:

- Discover optimal parameter settings for the 802.15.4 MAC protocol, which allow to maximize throughput (data rate) and minimize single-hop communication delay.
- Determine expected lifetime of a network with a given load traffic requirements.
- Achieve the objective of energy-efficiency and consequently extend the lifetime of a wireless sensor networks.

Upon completion of the model and having understood the limits of the standard regarding medium access we will propose modifications to the CSMA-CA algorithm. In the literature we can find several energy-efficient MAC protocols with duty-cycling mechanisms. All of these proposals are characterized by a high dependence in the application scenario and use of packet overhead to execute the mechanism. For instance T-MAC protocol [12] employs a future-request-to-send packet to overcome the early sleeping problem in a multihop communication.

Thus far, MAC protocols do not consider the current state of the receiving nodes. The sender always demands other nodes to be active in order to capture its packets. There exist cases in which either the energy level or the number of packets in the queue of the intermediate node determine that best action is to stay inactive. Therefore, the best strategy is to allow other neighbor nodes to receive, and hence retransmit the packet to destination. To find an appropriate energy-efficient mechanism for such cases is the ultimate objective of this research project. The initial proposal envisions a lifetime optimization of wireless sensor networks by means of sleep-wakeup adaptive mechanism in which each node decides to be active based on:

- Energy level of node.
- Number of packets in buffer and prioritization.
- Network delay requirements.

### III. METHODOLOGY

Analytical models are useful tools for evaluating the performance of MAC protocols, making easier the deployment and optimization of the network. The types of stochastic models for IEEE 802.15.4 MAC protocol which can be found in literature are: Markov-based, Renewal and Mean-average analysis.

Since Bianchi [13] Markov chains have been a widely used technique for modeling wireless networks (IEEE 802.11x). In this technique a Markovian model of the system is developed and its state transition probabilities need to be found. The state space of the model increases with both the complexity of the protocol studied and the number of users in the system. The authors of [15] and [18] presented Markovian models for the MAC protocol of IEEE 802.15.4 standard. Pollin et al. in [15] found a solution for saturated traffic conditions. Misic et al. presented simulation results in [18] which deviates from their own model.

Renewal analysis is a technique used in [14] which models the behaviour of a specific node of the network, named tagged node. The concept is based on a three-level renewal process which presents a fixed-point solution. The cycling/repetitive behaviour of a node provides the base for this technique. Although the renewal process in [14] works for saturation, it fails to model unsaturated traffic conditions.

In the first phase of the project we are building the analytical model for an unsaturated homogeneous network of N transmitting nodes, each reporting data via uplink to a receiver device through single-hop communication. We use a mean average technique, which has been previously utilized for other MAC protocols with exponential backoff. The analysis consists in finding the average duration of the backoff period of an individual node and the probability of success at the end of each backoff stage. The set of equations can be solved with numerical techniques.

The performance of the protocol depends on the incoming MAC frame traffic at every node’s queue. In general, two cases can be identified. If there always exit frames in every node’s queue we have a saturation condition. If frames are randomly generated at each node according to a Poisson process of certain rate λ, such that the tagged node has period of time with an empty queue after servicing its previous frame, is unsaturated condition.

Our work differentiates from existing analytical models by the fact that the queue size is not assumed to be equal to one. Inversely nodes have buffering capabilities as occurs in real applications with latest sensor transceivers [19][20]. Additionally the mean-average approach simplifies the mathematical analysis while keeping the same or higher accuracy of other approaches.

### IV. STATUS AND VALIDATION

We have recently succeeded in building the analytical model for a wireless medium which is error-free and without hidden and exposed terminal problems. There is no data acknowledgment and the CAP period is assumed to be infinite. This is a good approximation of a real scenario for an IEEE 802.15.4 network since the sink rarely communicates with nodes to pass control messages and the periodic beacon frame lasts very few slots, i.e. generally 1.8 slots.

We found the probability that a node attempts to sense the channel in any given slot (sensing rate φ) and also the likelihood of failing an attempt (failure probability γ). The analysis has been divided in 2 parts: the backoff process and slot state, and produces a series of equations for φ and γ that were solved for the IEEE 802.15.4 standard parameters by using numerical techniques. With these probabilities, the most notable performance metrics normalized throughput, MAC inter-successful transmission and average power consumption time can be evaluated.

Simulations can be conducted to validate the accuracy of the analytical model. Two alternatives are available: create our own simulation platform based on a pseudo CSMA-CA algorithm or use an existing network simulator. Opnet Modeler is a commercial tool for testing network designs in realistic scenarios and offers a built-in ZigBee/IEEE 802.15.4...
feature. NS-2 is an open-source network modeler with higher acceptance among researcher because it provides substantial support for simulations of TCP/UDP, routing and multicast protocols over wired and wireless networks.

For our current work we have used the IEEE 802.15.4 WPAN module of NS-2 and set up a sensor network with star topology and beacon-enabled modality. With this parameter setting the performance of MAC sublayer is effectively evaluated. One of the early results obtained can be seen in Fig. 1. The network normalized throughput, which is a measure of the average data received in the sink in every slot, is depicted for a network of 20 nodes with a Poisson-arrival traffic in the range of $0.5 \leq \lambda \leq 25$ milliframes/slot. Finally, Fig. 1 proves the accuracy of our new model and Fig. 2 shows the analytical average power consumption of a node.

V. CONCLUSION

In this extended abstract we reviewed the aims of research of our project on lifetime optimization of wireless sensor networks. The problem, motivation, current methodology and approaches used thus far were explained. We briefly described the progress in building a analytical model for IEEE 802.15.4 MAC protocol for unsaturated conditions. Note that our study differentiates from existing works on analytical models because we use real traffic scenarios and buffering in nodes. These results represent initial contributions for the lifetime optimization objective. With this model we can find several optimal parameter settings for the 802.15.4 MAC protocol, which maximize throughput and minimize delay, of a wireless sensor network.

In the following months we will focus on modeling a lossy wireless medium, where packets could be lost because of communication propagation effects and interference. This way the first phase of the project will terminate. Then, we will use the acquired knowledge to propose enhancements and/or modifications for the medium access mechanism of WSN.

REFERENCES

Specification, Integration and Management of Security in Service-Oriented Software Systems

Tan Phan
Faculty of ICT, Swinburne University of Technology, Melbourne, Australia
P.O. Box 218, Hawthorn, VIC 3122, AUSTRALIA
{tphan}@swin.edu.au

Abstract—Ensuring business-oriented security requirements are property interpreted, operationalized during development time, and enforced at runtime is a major challenge that current research has not addressed in full. In this paper, we discuss the HOPE framework introduced in the author’s PHD thesis. HOPE supports in a systematic manner the specification of security-oriented policies at the business level and their refinement into security policies at the system/service level. HOPE is also aimed at defining an effective mechanism for verifying services against security requirements and discovering services with compatible security settings. Finally, a mechanism is defined for HOPE that, at runtime, enables the protection of sensitive data that might flow into components not known at design time.

Keywords: security requirements, security objective, security verification, security-based discovery, data protection

I. RESEARCH PROBLEMS AND OBJECTIVES

The heterogeneity, dynamism and business-oriented nature of SOA make addressing security in service-based systems a challenging issue. In developing an SOA application, there are essentially three types of services/components 1) new components that are developed from the beginning 2) known legacy or third party components which are integrated into the system at design time or deployment time, and 3) unknown components that the system dynamically engaged through message exchange at runtime such as third party services or clients. Each type of components requires a different approach for security development and enforcement.

For the development of new components the main issue is in ensuring the “requirement traceability” for the development process. That is to guarantee that the components being developed fully realize all the security requirements. In current development practices, high-level security requirements are often identified by business analysts while the realizations of such requirements are implemented by engineers and developers. Linking the high-level requirements (specifying what need to guaranteed) and the realization mechanisms (specifying how such guarantees are realized) is a difficult task. This, therefore, creates the need for a systematic process and related techniques that can derive the system-level realization from the business-level requirements.

In contrast to new component, third party or legacy services that are integrated with the system at design or deployment time either exists before the development of the SOA project or is provided by third party for which the organization does not have control over. Therefore, even with the existence of a systematic process to derive the needed system-level mechanisms from the business requirements, it cannot be assumed that such legacy and third party components were developed following such process and thus offer all the required security requirements. Therefore, there is the need for a mechanism to ensure any given legacy or third party service has the necessary realization mechanisms to meet a system’s quality objectives before being used.

For services unknown at development time that are dynamically engaged with the system at runtime, ensuring quality compliance for them is much harder. SOA systems involving third party components are essentially multi-stakeholder distributed systems. In such a system, each stakeholder needs to ensure their requirements are met by previously unknown third party components that are engaged with their system. Such requirements need a dynamic mechanism at runtime to be propagated, analysed and enforced. At the moment there is the lack of a mechanism that enables such requirements to be propagated to the destined services and being analysed and enforced by them.

The objective of this PhD research project is to provide an extensible framework addressing problems identified above. As a result of the author’s PhD work, we propose the HOPE (High-level Objective-based Policy for Enterprises) framework which is aimed to allow for
- Business-level specification of security objectives and their refinement into system-level realizations
- Service quality conformance and compatibility checking during registration and discovery
- Runtime propagation of security requirements to the services/components unknown at development time

II. RELATED WORK AND THEIR SHORTCOMINGS

We have reviewed existing work in developing and managing secure systems and assessed their application for SOA systems. We have identified some shortcomings of existing work.
First of all, there is inadequate support for business-oriented security requirements specification and the lack of a systematic security-engineering process for SOA systems. There exists a number of policy frameworks and languages such as PONDER [1]. However, they fall short in enabling the specification of business requirements and the refinement of them into system-level policies (for more details see [2]). There have been a number of attempts to apply model-driven architecture (MDA) techniques for the modelling and translation of security attributes into system-level realization mechanisms such as [3]. However the entities being modelled in such work, even though being platform-independent, are still technical entities representing technical concepts such as filter, connector, services, and proxies. This means the models are technical and complicated which prohibit the participation of business practitioners in the modelling process.

We have also identified that there is inadequate support for discovering services based on security requirements. There have been a number of research efforts in enhancing the capabilities of services discovery by incorporating non-functional aspects, including security into the discovery criteria. For example, UDDI [4] attempts to extend UDDI with the notion of ‘blue pages’ for enabling service discovery based on user-defined properties like Quality of Service (QoS) that a service can provide, or the methods available within a service. However, security is not considered in details in UDDI. In particular, how security properties are modeled in the queries or represented in the registry data model are not mentioned.

Finally, the issue of runtime data protection in multi-stake holder systems have not been addressed in full. There have been a number of research efforts on data flow protection for SOA, particularly for service composition. Work such as the SCIFC model in [5] addresses the problem of access control for sensitive information when the information flows in a chain of Web Services. While preventing data flowing to undesired services is important, it is equally important to ensure that any service, upon receiving the data, have to protect the data according to the owner’s or generator’s requirements. SCIFC and other work in information flow security have not addressed that.

### III. THE HOPE FRAMEWORK

![Diagram of the HOPE framework](image)

In this section, we introduce the HOPE (High-level Objective-based Policies for Enterprise) framework (Figure 1) which is designed to achieve our research objectives. HOPE focuses on providing traceability for the security development process, automation of the security verification and discovery, and dynamism in the propagation of security requirements. HOPE comprises three main elements. First of all, HOPE includes a systematic step-wise process (SOABSE) and related models and techniques for specifying security objectives and translating them into system-level realization mechanisms. Such systematic process ensures security requirements are traceable in development. Secondly, HOPE contains a registry-centric (REGISTRY) set of models and techniques for verifying that certain services are compatible with specified security objectives and identifying matching services given certain desired quality objectives. Such models and techniques allows for the automatic verification of given service’s security settings against the requirements and automatic filtering of services with incompatible security offering. Finally, HOPE includes a set of techniques for the propagation (PROPAGATION ENGINE) of security requirements, following the message flow, to services that the system interacts with at runtime. Such techniques ensure that requirements are dynamically propagated without the need for a centralized system controller or a known system topology before hand.

#### A. SOABSE: The Security Engineering Process

The first major component of HOPE is the SOABSE (SOA- Business Security Engineering) process and related techniques to improve the current security development practice for WS-based software systems [6]. In SOABSE, security objectives are modeled at business-level which are then systematically mapped into security measures performed on WS in the form of WS security policies. SOABSE includes a generic security model which maps security attributes and their realization mechanisms and a meta-model for capturing application-specific security deployment information. A set of transformation procedures that help automate the translation of security objectives into system-level security functions performed on Web Services and their elements has also been developed.

**Generalization of SOABSE for other Service Qualities:** The SOABSE process has been extended to accommodate for other system properties such as reliable messaging [7]. The generalized process allows the systematic specification of quality-oriented policies at the business level and their refinement into policies at the system/service level. Quality-oriented business requirements (quality requirements) are expressed as quality objectives applied to business entities which are modeled in application entity model. These objectives are then refined or translated into system-level WS-Policy statements. The refinement relies on an application-specific business entity model and application-independent domain quality models.
B. REGISTRY for Security Verification and Discovery

The second major component of HOPE is Registry-centric set of models and techniques to address the issue of security-oriented service registration and discovery [8]. A service registry has been proposed that serves as policy storage and management facility, a policy checkpoint during service publication and as a policy matchmaker during service discovery. We have also extended WS-Policy with a policy conformance operator for policy verification at service publication time and used WS-Policy Intersection for policy matching at service discovery time. A policy information model and policy processing logics are also developed, which are encoded in a Policy Validator and Policy-Enabled Query Manager components of the registry.

Extension of the Registry for Trust-based Discovery: The REGISTRY [8] has been extended for effective trust-based service discovery [9]. This is a collaborative work with NICTA. Techniques proposed in [8] is employed to integrate policy requirements with trust mechanisms for both requesters and requested services. In particular, the extended registry component in [8] is used to store the trust-based policies and match requested services and requesters’ policy requirements.

C. Runtime Data Protection Requirements Propagation

The last major component of HOPE is PROPAGATION ENGINE, a set of techniques for runtime dynamic propagation of protection requirements [10]. With such engine, we propose an approach to aid collaborative partner services in properly protecting each other’s data. Our approach allows each partner to derive an adequate protection mechanism for each message it sends based on those of the corresponding messages it receives. We modify the message handling mechanisms of Web Services engines to dynamically gather protection requirements for a given outgoing message as an aggregation of requirements from original owners of data in the message.

IV. VALIDATION, PROGRESS AND REMAINING WORK

The work related to the HOPE framework and its component have been finalized as discussed. A case study with realistic business scenario and a comprehensive set of rules and regulations has been developed to validate the HOPE framework. In the case study, security-oriented requirements from Australian Privacy Act were extracted to derive the quality-objectives. The case study then validates elements of the HOPE framework (including the process, its models and techniques) against such derived quality objectives to verify the comprehensiveness and practicality of HOPE. Main components of the HOPE framework have been implemented and validated using the case study. The project is in its final stage and the author is finalizing his thesis.

V. RESEARCH CONTRIBUTIONS AND CONCLUSIONS

We have presented in this paper an outline of the research involved in the author’s PhD thesis. The main contribution of the thesis is the HOPE framework. HOPE is created for aligning business-oriented rules and requirements with system-level management via a mechanism that allows for the specification of security-oriented business rules and regulations and the refinement of them into system-level security mechanisms. With HOPE, we also provided a general mechanism that utilizes a service registry for security-based service registration and discovery. HOPE also provided a set of techniques for the dynamic propagation of security requirements to services that the system interacts with at runtime following the message flow.

ACKNOWLEDGEMENT

The author would like to thank Professor Jun Han, Doctor Jean-Guy Schneider, Doctor Steve Versteeg, Doctor Ingo Mueller and many others for their guidance and support in this PhD thesis.

REFERENCES

Multi-agent System for Local Data Propagation in City Traffic Monitoring
Integration of Modal and Fuzzy Methods for Agent’s Knowledge Representation

Grzegorz Popek
Centre for Complex Software Systems and Services
Swinburne University of Technology
Melbourne, Australia
gpopek@swin.edu.au

Grzegorz Popek
Institute of Informatics
Wrocław University of Technology
Wrocław, Poland
grzegorz.popek@pwr.wroc.pl

Abstract—An aim of this research is to propose a multi-agent system capable of local city traffic monitoring, including: summarization of collected data, local propagation of data in human-understandable way (using linguistic terms), aggregation of messages from adjacent agents, and usage of modal operators (of possibility, belief and knowledge) to model uncertainty of a summary related to lack of data from sensors. A three-layer model of a system used to model single agent is briefly described.

data aggregation; local data propagation; multi-agent system; linguistic summarization; knowledge integration

I. INTRODUCTION

Solving problems in a system at a global level can lead to unacceptable complexity; especially in systems designed to work in a real time. One of possible solutions is a local approach. That is, instead of solving a complete problem at a central level, a locally-optimal model can be constructed (see [9]) by decomposing main problem into sub-problems related to separate elements of the system. Often locally-optimal model is outperformed by a globally-optimal model. Still, this is an accepted tradeoff for a reduction of complexity similar to one found in heuristic approach.

Multi-agent systems approach to local problem solving models a system with multiple independent autonomous objects capable of communication (agents) which are assigned with different local problems, modeled with an use of their preferences. The goal of each agent is to satisfy its preferences – which should be understood as a step toward a local optimization of a model – with a global preservation of constraints between sub-problems. As a result, one expects an emerging behavior of such a distributed system giving a good global performance.

In proposed research, the main goal of modeled system is to provide accurate and up-to-date information about city traffic load to an external user. An aim of this research is to propose a protocol for local propagation of gathered information in a transparent way. A system is assumed in a fully distributed way as a net of agents equipped with traffic sensors and equal roles. An achievement of a wanted level of transparency is planned through modeling of propagation (inter-agent communication) with linguistic formulas (fuzzy-logic statements – compare [7,8,10]).

Also, a situation in which certain part of data is missing (which can happen both at a sensor level and at inter-agent communication level) should find its reflection in summarizing statement. An idea is to model such situations using epistemic operators of Possibility, Belief and Knowledge.

Last part is to define a way of integration of fuzzy statements coming from surrounding agents (possibly with modal operators) and relate it to agent’s private sensory input to obtain a consistent summarization (compare results for integration with crisp language and graded perceptions in [4,6]).

II. APPLICATION SCENARIO

Let a network of roads represented as a directed multi-graph be assumed. Each crossroad is represented as a unique vertex and each one-way connection between two adjacent vertices is represented as an arc. It is of course also possible to represent only a few points of a whole road-network as vertices and describing possible ways between them as arcs. As for this research, the first representation will be used. It allows such assumptions as uniform traffic load within each separate edge. Of course, such a multi-graph can be represented as a graph by introducing additional vertices into multiple arcs connecting same pair of vertices. Still, assumed representation is not an issue in this paper, so let this problem be left unattended for now.

Let it be also assumed, that with each edge a capacity and length (i.e. travel cost) is correlated. Also, that a monitoring unit capable of gathering traffic load data is located at each edge.

It is also assumed that there is a set of users travelling across the graph. The goal is to provide a user with up-to-date and accurate information about current traffic state around her location. The idea is to make it through local announcements of each monitoring unit. Also, assuming that monitoring units (agents) also can receive messages from other ones, how to aggregate them to make a statement about bigger area.

III. MODEL OF SINGLE AGENT

Let a 3-layered model be assumed, where bottom layer represents data, middle layer represents concepts and top layer represents language symbols.

The main idea is to use formulas of concept-based language defined at a top level to reflect a current state of world described in a database. In case of complete
knowledge about a current state of an environment (when there is no missing or unobserved data), data can be directly correlated with concepts and summarization can be determined. There are many problems which need to be solved. The first one is connected to a fact, that data stored in a database can be partially missing. It is important to define a way, how to relate a concept to an incomplete tuple from a database.

Another one is how to relate concepts to each other – not only with a distance function by also with an application of linguistic hedges (also called linguistic modifiers). Based on a proposed solution an embodied thesauri will be built.

Last but not least important problems are met at the highest level. There is a need to generate a human-understandable summarization of a database with a language based on concepts. An interesting case is met, when databases are indexed with time and are interpreted as an agent’s experience base. In such situation there is a need to generate a message about a concept which cannot be connected to the actual state because of a lack of respective data, it can be successfully represented with an application of modal operators and message can be formulated based on previously gathered data.

A. Data Layer

Data layer can consist of empirical observations of an artificial agent, of some financial transactions’ data stored in databases, or of history-data about types of interaction between users in social network. In proposed scenario, it is data about a traffic load within observed area.

It is assumed, that multiple data sets consist of sets of ordered \( m \)-tuples from the universe \( U = [0,1]^m \). For each tuple from each \( z \)-th data base, it is indexed with a time index \( t \) and is marked as \( W(t) = (w_{1}(t), w_{2}(t), \ldots, w_{m}(t)) \in U \). This model is quite flexible as allows many interpretations. It can be assumed, that for any but set \( z \) and for every \( i \in \{1,2,\ldots,m\} \), every \( w_{zi} \) is independent and reflects separate dimensions of a description. Still, we can treat each \( w_{z} \) as data coming from different sensors, where some of sensors can be correlated with the same dimension (e.g. three sensors of different lengths of waves of light; two types of transaction excluding each other, traffic load in two subareas of a same road).

Although this representation clearly reflects only unary relations, binary relations (e.g. between users) can be reflected by representing an interaction of each user with others with an individual data set of such a form.

B. Concept Layer

Middle layer represents concepts constructed over the space of data. It is assumed that an agent is equipped with a set of concepts. As for an example, concepts can be given by an expert to represent certain interesting configurations of data values in database. In case of social network approach concepts can be treated as some communication patterns (e.g. work relationship – regular exchange of documents; boss relationship – one-way communication consisting of information messages; friend relationship – frequent communication and exchange of multimedia content).

Concepts are represented as fuzzy variables over the universe \( U \). Therefore, each concept \( C_{\mu} : U \to [0,1] \) is represented by a membership function \( \mu_{\mu} \mid U \to [0,1] \). Also, it can be assumed that mentioned membership function \( \mu_{\mu} \) can be derived from a set of concept representatives \( C_{\mu} = C_{\mu(1)}, C_{\mu(2)}, \ldots, U \). Although, in this case construction of a similarity function which reflects properties of the universe \( U \) is inevitable. Given such a similarity function \( \sigma \), it can be assumed (as an example), that

\[
\mu_{\mu}(u) = \max_{\mu(1),\ldots,\mu(j)} \sigma(u, C_{\mu(1)}, \ldots, C_{\mu(j)})
\]

for every \( u \in U \).

C. Language Layer

Top layer is a layer of language symbols. This layer offers a way to summarize states of collection of stored. It can define a language of agents’ communication, a language for a summarization of a database or a language used to inform an administrator about types of relations between users of a social network.

There are two main types of messages planned for the system. First are non-modal messages. They are used mostly in a summarization of databases. A goal is to reflect a state of a database with human-understandable messages. The problem is the most interesting, when data stored in a database is incomplete, because this incompleteness can be treated in various ways by applying a priori knowledge.

Second types of formulas are modal formulas. They are used when we assume that a database is a set of agent’s observation. Therefore, the agent can refer to currently unknown properties of an environment based on its past experiences. It models its beliefs and confidence using modal operators.

IV. OPEN QUESTIONS

First open question is connected to low-level aggregation of data about a traffic state observed at a given edge. How to aggregate multiple values to make a fitting summary for this application?

Most of authors agree that (after [1]) for each natural \( n \geq 2 \), an operation \( F : [0,1]^n \to [0,1] \) is called an aggregation function (\( n \)-ary aggregation operator) if it is increasing with respect to each variable and fulfills boundary conditions \( F(0,0,\ldots,0)=0 \) and \( F(1,1,\ldots,1)=1 \).

This flexible definition allows for a wide choice of operators (e.g. quasi-arithmetic means, triangular norms, triangular conorms). Still, it is difficult to justify why one of them is better than other ones other than analyzing preservation of some commonsense requirements.
Therefore, to answer a question about a choice of an aggregation function, an analysis of its behavior in exemplary cases is needed. Also, assuming a set of commonsense requirements, a proof of their preservation is needed.

Aggregation at a higher level yields other problems. How to integrate messages from other agents? And what (although for now there is an assumption, that agents are uniquely assigned to graph’s edges) if there are multiple agents observing the same road getting inconsistent results? This may lead to problems of reaching consensus and to a problem of belief fusion. There are many approaches to this problem (see [3,5]) including such as: additive, non-additive, qualitative beliefs.

Most of mentioned approaches leave to some paradoxes. For example, lottery paradox. Imagine a situation, where agent believes in two statements. Should it also believe in their conjunction? It would be nice. But now, imagine a lottery with 100 tickets. Each of those tickets gives an equal chance of winning. An agent gets one ticket, and since a chance to win is so low, due to some model (with a certain threshold for holding beliefs) it believes that it will lose. And that is true for every tickets. It believes that every ticket separately is a losing one. If an agent was able to believe in a conjunction of beliefs, even having all of tickets, it would still believe in losing lottery.

Finally, how to build a language of summaries and how complex it has to be to reflect properties of load in surrounding area. Also, how to propagate information about load between agents. Is it needed to include into summaries information about load at areas placed in different directions from an agent and how big area needs to be covered. Such descriptions will be built with a use of fuzzy quantifiers. But it carries some additional problems.

Linguistic quantifiers are often based on fuzzy cardinalities (see [2]). It is often stated, that cardinalities of fuzzy sets should be represented as fuzzy numbers for example through α-cuts. But communicating whole fuzzy number can take a lot of bandwidth. Still, representing cardinality of fuzzy set leads to some unwanted behaviors. Cut-based approach leads to defuzzification and therefore whole aspect of fuzzy approach is abandoned through it. On the other hand, mass based approach gives unwanted properties. Classical example of such a property is to assume a set of 3 men $\mathcal{X}=$\{John, Sam, Ted\} and a fuzzy property \textit{tall}. Also, let $\mu_{\text{tall}}(\text{John})=1$, $\mu_{\text{tall}}(\text{Sam})=0.5$, $\mu_{\text{tall}}(\text{Ted})=0.5$. In mass-based approach, cardinality of set $\mathcal{X}$ would be $1+0.5+0.5=2$. But in reality? Since $\mu_{\text{tall}}(\text{John})=1$, he needs to be assumed to be a tall person. What about the other two? If Sam is assumed as tall, Ted also should be, since $\mu_{\text{tall}}(\text{Sam})=\mu_{\text{tall}}(\text{Ted})$. Therefore, cardinality of set of tall people would be either 1 or 3, but never 2.

It leads to many propositions of convex and non-convex cardinalities. There are also ideas of applying only partial orders to express cardinalities of families fuzzy sets over a certain domain. Unfortunately, it does not really enable system designers to use such mechanisms to calculate something useful. Of course, one should not neglect a need for formalizations and analyses of properties of families of fuzzy cardinalities in general. Still, such analyses are far away from applicability in a real environment.

REFERENCES


SeRenDiP: Towards Service Relationship Driven Processes

Malinda Kapuruge
Faculty of Information & Communication Technologies
Swinburne University of Technology, P.O. Box 218, Hawthorn, VIC 3122, Australia
mkapuruge@swin.edu.au

Abstract—Traditional business process modeling approaches for service compositions do not explicitly represent the service relationships. The mutual behaviors, responsibilities are tightly bound to the orchestration and intermingled with the other control and data flow logic. Thus, the modifications to face environmental disturbances such as unavailability, variability of services are tedious. In our approach we try to model business processes based on explicitly defined service relationships. Such explicitly defined service relationships form an organizational structure that regulate and reconfigure itself to face the unexpected changes at runtime.

Keywords—component; BPM, Flexibility, Service Relationships, Service Composition

I. INTRODUCTION

Service compositions such as service brokers, mediators, emerging to compose asynchronous, loosely coupled services offered by different vendors[1]. The collective behavior of these service providers is used to achieve the business goals of the composition. Therefore coordinating the offerings of such variety of participants is important. Thus, providing automated support for business processes in Service Compositions is critical to coordinate the offerings of different service providers and requirements of consumers.

However such an automated support should not unnecessarily restrict the composition from adjusting to the market conditions to capture new business opportunities. In traditional business process modeling mechanizations the concrete services and their expected behavior is tightly bound to the process specifications. As an example a WS-BPEL script acts like glue between the existing web services, orchestrating the service invokes. However in such a setting, survival of business process instances gets challenged due to many causes. This includes the unavailability of services, the changes in dependent service’s functional behavior, etc. Thus business process definitions need to be evolved and the process instances need to survive the exceptional situation.

In the past many approaches attempted to address the above mentioned limitations of existing standards such as BPEL that are designed for business process modeling in services compositions[1-4]. Instead some attempted to find alternative paradigms [5-9].

However, an evaluative survey carried out by us on past approaches, showed that the lack of representation of inter service relationships in services compositions cater for inflexibilities in business process modeling[10]. Therefore our attempt is to improve the flexibility in process aware service compositions, via an explicit representation of inter service relationships.

II. SERVICE RELATIONSHIPS

A service relationship defines the expected mutual behavior among two collaborators in an abstract manner. As an example assume a road side assistance business as shown in figure 1. The service broker (business) may define certain terms and responsibilities under which a service provider (Garage or Tow car) and a consumer (Client) should collaborate. E.g. The payment should be done prior to the repair. At runtime different Garages, Tow cars and Clients can play these roles by maintaining defined relationships. e.g. Mr. Smith, Tom’s repair, EZ-Tow as in figure 1. However, a certain Garage might accept only the credit card payments. Therefore the service relationship gets more concrete with that new knowledge at runtime and the process model should be open for such behavior specializations. Although this could be achieved via a “switch” in a re-defined workflow, the unpredictability, complexity and variability of service behaviors might make continuing modifications completely impractical.

Also there are other advantages of explicitly representing the service relationships. First, without such it is difficult represent the knowledge of two participants in terms of their collaboration aspects. Such collaborations are common in automated service oriented computing environments and are usually evolved overtime. Second, it is possible to derive the behavior of a participant based on its relationships with others. In other words, what defines the behavior of a Character/Position/Role is the projection of its behavioral relationships with the rest as shown in figure 1. Such a representation of Role is important in business process modeling [11]. Third, such an explicit representation of service relationships provides the basis to define the business processes of the composition. In the sense, a business process defined to achieve a particular goal is realized by the well defined relationships of the participants. Fourth, as both
the roles and the business processes are based on well defined relationships, the changes in those relationships are automatically get reflected upon them.

III. PROCESSES

The representation of such service relationships should be done in a manner, so that the composition to maintain a homeostatic and a morphostatic association with its operating environment[12]. This homeostatic and a morphostatic properties of the composition are achieved via regulation and reconfiguration of service relationships. As the processes are specified based on such regulatory and reconfigurable service relationships, such changes automatically get reflected upon processes. Further, a service composition maintains both the long term and short term business goals. The strategies to achieve long term business goals are usually persistent compared to the short term goals and usually result in evolutionary changes to the composition. However, in order to survive in highly volatile business environments, the composition must also support short term strategies as well. Usually these strategies are formed to handle exceptions or to quickly grab a business opportunity via a hot fix. Most of the time the scope is limited to a single process instance. Therefore the business process models for service composition should improve the flexibility in both process types and process instances.

IV. APPROACH

ROAD framework [13] define an organization as a collection of inter service relationships. In our work we further improve this to define expected behavior between two collaborating parties. Each service relationship defines a number of behavior terms. A business process specification in SeRenDiP is a logical group of such behaviors. Therefore changes in service relationship behaviors automatically get reflected in processes.

![Figure 2: Processes based on service relationships](image)

Adaptation could be carried out both via regulation and reconfiguration of the organizational structure. Regulations include the modifications in event patterns, obligatory roles, post events etc. Reconfigurations include duplication of service relationships, redefinition of business processes, process merging/de-merging etc. Regulatory changes are automatically getting reflected in business processes, whilst the reconfigurations might require modifications to business process definitions.

Also the adaptation could be differentiating according to the persistence of change, i.e. Evolutionary and the instance specific. Evolutionary adaption is carried out at the process type level. Instance specific adaption is carried out only to fix an issue in a process instance[14]. One of the great advantages of the SeRenDiP framework is its ability to provide reconfiguration of the structure for both the evolutionary and the instance specific changes. As an example, let’s say for the scenario shown in figure 1, there is requirement to add another Garage service to speed up the repair for a client, the framework provide the feature to duplicate the Garage-Client relationships.

However, the biggest challenge actually is not providing the flexibility but to preserve the business invariants upon such flexibility. The flexibility should not come at a price of compromising the business invariants. Thus in our approach we have developed certain guard conditions to ensure the soundness of the modifications. This includes avoiding the live locks, deadlocks, partial executions, cyclic executions of processes.

V. RELATED WORK

In the past many approaches were taken to model and enact adaptable business processes. Some approaches were specifically designed for service compositions [1-3, 8, 15-17], whilst some are generic [6, 18-20]. In another angle, some approaches could be seen as improvements for existing modeling standards [1, 2, 15, 21], whilst others could be seen as alternatives [6-9, 22].

In [15] authors provide an approach for autonomic handling of service failures in WS-BPEL[23], which they improve in [1] for dynamic service selection. However it is not possible to alter the internal behavior of the composition. Hence the flexibility is limited to dynamic service selection for an already expressed business process. Also there is no way to treat each process instance uniquely. Charfi et al. in [18, 21] propose an aspect oriented approach to manage the adaptation. The solution reduces the complexity of runtime adaptation. However the adaptation is limited to defined point cuts. Also it is not possible to specialize a process instances to the runtime as a change affects all the process instances.

The importance representation of Role in BPM is highlighted by Balabko and Wegmann in [11]. Saidani and Nurcan in [24] attempts to model flexible business processes based on Role, Mission and Operation concepts. Further, role Interaction Nets[25]. Role Activity Diagrams[26] use the notion of swim lane to group a set of activities in a business process to a single responsible participant. However, rather than defining the behavior in terms of Roles, our approach suggests to derive the behavior from the adjoining service relationships. Consequently, improving the maintainability and flexibility of a composition.

VI. PROGRESS AND FUTURE WORK

Existing standards for business process modeling has been reviewed. Later those were analyzed against a criterion defined by us to identify the flexibility limitations. Based on those identified limitations, possible solutions were analyzed[10]. A process meta-model has been developed. A language has been defined to specify the relationships behaviors and business processes. Syntax has been developed to specify the behaviors and is under refinements. A case study has been developed based on a service broker
that provides road side assistance to a set of registered clients in Australia. Case study is used to validate/refine the meta-model.

As future work, the meta-model will be validated against similar case studies. Later the meta-model and the API will be further improved to delegate the management among different participants of the composition.

ACKNOWLEDGEMENT

Many thanks to my supervisors; Prof. Jun Han, Dr. Alan Colman and Dr. Minh Tran for their generous guidance.

REFERENCES

Answering XPath Queries Using XPath Views

Rui Zhou
Faculty of Information and Communication Technologies
Swinburne University of Technology
Melbourne, VIC 3122, Australia
{rzhou}@swin.edu.au

Abstract—Answering queries using views (also known as rewriting queries using views) is to utilize previously defined (possibly materialized) views to evaluate queries in order to save the cost of accessing large real database or provide a privacy-preserving publishing. It is a classic problem, and appears in many applications, such as query optimization, data integration, data warehouse and query caching.

With the prevalence of XML technologies on the web, rewriting XML queries using XML views has caught the attention of both researchers and system designers, and is believed to be a promising technique in web application development. Since XPath serves as the core sub-language of the major XML query languages such as XQuery and XSLT, we focus on answering XPath queries using XPath views.

I. APPLICATION BACKGROUND

The first class of applications in which we encounter the problem of answering queries using views is query optimization and database design. In the context of query optimization, computing a query using previously materialized views can speed up query processing because part of the computation necessary for the query may have already been done while computing the views. Such savings are especially significant in decision support applications when the views and queries contain grouping and aggregation. Furthermore, in some cases, certain index can be modeled as precomputed views and deciding which index to use requires a solution to the query rewriting problem. In the context of database design, view definitions provide a mechanism for supporting the independence of the physical view of the data and its logical view. This independence enables us to modify the storage schema of the data (i.e., the physical view) without changing its logical schema, and to model more complex types of index. Hence, several authors describe the storage schema as a set of views over the logical schema. Given these descriptions of the storage, the problem of computing a query execution plan (which, of course, must access the physical storage) involves figuring out how to use the views to answer the query.

A second class of applications in which our problem arises is data integration. Data integration systems provide a uniform query interface to a multitude of autonomous data sources, which may reside within an enterprise or on the World Wide Web. Data integration systems free the user from having to locate sources relevant to a query, interact with each one in isolation, and manually combine data from multiple sources. Users of data integration systems do not pose queries in terms of the schemas in which the data is stored, but rather in terms of a mediated schema. The mediated schema is designed for a specific data integration application, and contains the salient aspects of the domain under consideration. The tuples of the mediated schema are not actually stored in the data integration system. Instead, the system includes a set of source descriptions that provide semantic mappings between the relations in the source schemas and the relations in the mediated schema. Most data integration systems follow an approach in which the contents of the sources are described as views over the mediated schema. As a result, the problem of reformulating a user query, posed over the mediated schema, into a query that refers directly to the source schemas becomes the problem of answering queries using views.

In the area of data warehouse design we need to choose a set of views (and indexes on the views) to materialize in the warehouse. Similarly, in web site design, the performance of a web site can be significantly improved by choosing a set of views to materialize. In both of these problems, the first step in determining the utility of a choice of views is to ensure that the views are sufficient for answering the queries we expect to receive over the data warehouse or the web site. This problem, again, translates into the view rewriting problem.

Finally, answering queries using views plays a key role in developing methods for semantic data caching in client-server systems. In these works, the data cached at the client is modeled semantically as a set of queries, rather than at the physical level as a set of data pages or tuples. Hence, deciding which data needs to be shipped from the server in order to answer a given query requires an analysis of which parts of the query can be answered by the cached views.

II. RELATED WORKS

Answering queries using views has been extensively studies for a long time. Halevy [1] did a survey on this problem over relational database and pointed out its wide impact on a number of data management applications. Efficient algorithms were developed as well. eg. MiniCon [2], bucket [3], inverse-rules [4], [5], to tackle the problem in relational context.

It then immediately started to draw the attention of researchers on XML data. Since XPath serves as the core sub-language of the major XML query languages such as XQuery and XSLT, fruitful achievements have been made on rewriting XPath Queries with XPath Views. Two types of rewritings for XPath queries have been studied in the literature. One is equivalent rewriting [6]: Given a materialized view \( V \) of a
database $D$, an equivalent rewriting $Q'$ of a query $Q$, runs over the view $V$ producing the same set of answers as evaluating $Q$ over $D$, i.e., $Q'(V) = Q(D)$. Here, we use $Q(V)$ and $Q(D)$ to denote the returned query results by evaluating $Q$ on $V$ and $D$ respectively. However, an equivalent rewriting may not always exist, and moreover part of answers covered by the view are still valuable, e.g. in data integration scenario, the data sources are limited to cover the domain. It is very common that we cannot find a equivalent rewriting for a query. Therefore contained rewriting [7] is introduced and can be described as follows: Given a view $V$ on a database $D$, a contained rewriting $Q'$ of a query $Q$, runs over $V$ producing a subset of answers as evaluating $Q$ over $D$, i.e., $Q'(V) \subseteq Q(D)$.

Lakshmanan et al. [7] also proposed the maximal contained rewriting (MCR), which is a set of contained rewritings (CRs) and no other CR set produces more answers than the MCR does. But an MCR may contain redundant CRs, i.e. CRs that are contained in other CRs, which means answers produced by redundant CRs can be covered by other CRs in the MCR set. Obviously, it is unnecessary to evaluate the redundant CRs. We call an MCR with no redundant CRs an IMCR (irredundant maximal contained rewriting). Our work mainly focuses on the IMCR.

Containment for a fragment of XPath queries $XP[[]]\{/\}$, including branches, wild cards and descendant axes, is shown to be conNP-complete in [8], though for three subclasses (combining any two of the three features), $XP[[]]\{/\}$ and $XP[\{\}]/\}$, the containment problem is in PTIME. [8] also proposed a PTIME-efficient but incomplete algorithm to determine containment in $XP[[]]/\}$. And this homomorphism-based algorithm was there afterward extended or utilized by the works [6], [9] to evaluate equivalent rewritings of XPath queries using materialized views. The differences between their works and ours are: (1) they focus on equivalent rewriting while we focus on contained rewriting; (2) they focus on the aspect of computing complexity using queries as input, while we focus on designing efficient algorithm running on data input. Containment for XPath queries under DTD constraints, with disjunctions and variables can be found in [10], [11]. A high-leveled summarization can be found in [12]. Recently, the complexity of query containment in expressive fragments of XPath 2.0 has been discussed [13].

III. RESEARCH METHODOLOGY

How to find the irredundant maximal contained rewriting?

The IMCR is expressed as a union of a number of irredundant CRs. Previous method of finding these irredundant CRs includes two steps: (1) generate all the CRs, then (2) remove the redundant ones. In the first step, the goal can be achieved by finding all the useful embeddings, because each CR is associated with a useful embedding. In step (2), we need to further check query containment for these candidate CRs. However, taking into account that the number of CRs may be exponential, step (2) is rather expensive, because, even though we ignore the possible case that one CR can be contained in a union of some others, we still need to check query containment for an exponential number of query pairs (i.e. whether one CR is contained in another CR). In the worst case, although all the CRs are irredundant, we have to carry out costly but useless step (2) as a routine. Obviously, it will be more efficient, if we manage to find the irredundant CRs directly so that we can save the cost of refining the candidate CR set.

We first propose a technique called concise embedding to eliminate part of the redundant CRs. The key idea is: concise embedding tries to embed more nodes (but not as many as possible) and is a restricted form of useful embedding. We can prove that, for a query $q$ and a view $v$, a contained rewriting must be a redundant contained rewriting, if it is produced by a useful embedding $e$ from $q$ to $v$, but $e$ is not an concise embedding. We then develop an algorithm to find all the concise embeddings and the CRs produced by these embeddings.

By using concise embedding, we can eliminate part of redundant CRs. However, a CR produced by a concise embedding can still be redundant. This is due to the containment of component patterns. Firstly, a CR may not be minimized. And after minimizing a CR, another CR will be easily found out to be contained in the minimized CR. To minimize the CRs, an intuitive way is to generate all the CRs and then minimize each of them. This is correct but inefficient, because the number of CRs may be exponential, up to $2^{N_{ele}}$. Consider that the genuine reason of a CR pattern not being minimal is the existence of pattern containment between different component patterns, we could first determine containment between all the component patterns, and then avoid adding contained patterns so as to generate minimal CRs directly. The total number of component patterns is up to $2^{N_{ele}}$ and the comparisons between them is of complexity $O(N_{ele}^2)$. This is efficient in polynomial in contrast to in exponential of the intuitive method.

How to compute the irredundant maximal contained rewriting on materialized views?

We propose a polynomial algorithm to answer an exponential number of queries. From the previous section, we know an IMCR may consists of up to $2^{N_{ele}f}$ irredundant CRs. To find out the IMCR, we need to enumerate an exponential number of irredundant CRs. However, evaluating the IMCR against materialized views is not that expensive. An assumed inefficient method is to firstly find out the IMCR set, then evaluate the possibly exponential irredundant CRs against views and union the final results. In fact, considering the characteristics of the irredundant CRs, we are able to produce the same answer result by only evaluating up to $2^{N_{ele}f}$ queries, a linear number of queries.

The exponential number of irredundant CRs attributes to two different embedding choices of one path according to concise embedding. Let the number of such path with (two embedding choices) be $n$, then the number of CRs will be $2^n$. For a path $p_i$, we use $B_i$ to represent that $p_i$ is embedded by one case, and $B_i$ to represent $p_i$ is embedded by the alternative case. Then any irredundant CR is encoded with a boolean...
expression, like \( B_1 \lor B_2 \lor \ldots \lor B_n \). The truth assignment of the expressions shows how to generate the rewritings. Not surprisingly, the \( 2^n \) boolean expressions correspond to the \( 2^n \) irredudant CRs. Here, we ignore the fully embedded paths, since they are already satisfied in the view and do not contribute to the rewritings. We also ignore the paths which have only one embedding choice, because the corresponding condition subtree must appear in the irredudant CRs and do not inflate the number of rewritings.

As we see from above, an exponential number of rewritings contain only \( 2n \) boolean literals. Each boolean literal is associated with a condition subtree that needs further testing in the view. \( B_i \) or \( \overline{B_i} \) appears in every irredudant CR, therefore evaluating all the irredudant CRs against the real view may result in repeated condition test of \( B_i \) or \( \overline{B_i} \). The idea arises from how to avoid the repeated computation. Technique details are omitted.

**How to filter unanswerable query rewritings using views?**

The motivation of this filtering work is that users may issue a large number of queries against a view \( V \). Although to test whether a query \( Q \) can be answered by \( V \) (i.e. whether there exists a rewriting for \( Q \) using \( V \) ) is P-TIME efficient (Actually this is the result for XPath subset \( X^P(//,//,\ldots) \), and complexity is even worse as coNP-hard for queries and views in \( X^P(//,//,\ldots,\ast) \)), with complexity \( O(|Q||V|) \) for equivalent rewriting \([6]\) and \( O(|Q||V|^2) \) for contained rewriting \([7]\), it is still of great importance if we can cheaply filter part of (as many as possible) unanswerable queries for \( V \). Consequently, a lot of computation cost will be saved. Fig. 1 shows the framework of evaluating queries using materialized view \( V \). The shaded filtering step is the focus of this work. Obviously, the filtering step should possess three properties: (i) It should not introduce false negatives, which means if we can find a rewriting for a query using \( V \), the query should not be rejected; (ii) It should be more efficient than computing a rewriting for the query, otherwise we would rather directly find a rewriting. (iii) It should be effective, which means the filtering should allow less false positives. In real applications, there may also be a large number of views, for instance, in data integration, many sources publish their views in Local-As-View architecture. Therefore filtering techniques play a significant role when we have to deal with plenty of queries and views.

we devise a set of \( O(|Q|) \) algorithms to filter queries. And we study the filtering for both equivalent rewriting and contained rewriting. The basic idea is to verify if the structural relationships in a query could be satisfied in a view, given that label preserving and structure preserving are the key conditions in discovering a homomorphism (for finding an equivalent rewriting) or a useful embedding (for finding a contained rewriting). We use index to capture the structural relationships in the view, and develop two algorithms for equivalent rewriting, i.e. Lazy Algorithm and Eager Algorithm. Moreover, Eager Algorithm can be modified to support contained rewriting. We first study all of the above for queries and views in subset \( X^P(//,//,\ldots) \), featuring child axes, descendant axes and branches in XPath, and then discuss the problem for \( X^P(//,//,\ldots,\ast) \) including wildcards.

**IV. PROCESS REPORT**

I’ve finished the aforementioned parts, and I’m writing my thesis. I plan to submit my thesis by the end of June. Any discussion is welcome off-line, since the short report cannot cover many technical details.

**ACKNOWLEDGMENT**

I appreciate the supervision of my mentor Chengfei Liu, and the discussion with my group colleagues.

**REFERENCES**

CORE: A Framework for Context-Aware Interactive Systems with COmprehensive REasoning Technique

Muhammad Ashad Kabir
Faculty of Information and Communication Technologies
Swinburne University of Technology
Hawthorn, VIC 3122, Australia
akabir@swin.edu.au

Abstract—Development of context-aware systems is inherently complex. Those systems gathered context information from variety of sources (e.g., physical/virtual sensors, computing devices, user’s profile, software/application, and so on) to infer situations (high level contexts). Based on situations and user preferences the systems might interact with each other on behalf of user. The diversity of the sources of context information and the characteristics of the computing devices strongly impact the Quality of Context (QoC) information as well as the situations. Moreover, the context information is dynamic, temporal, and imperfect by nature. Therefore, a formal representation of context data within a model is necessary for consistency checking, as well as to ensure that sound reasoning is performed on context data. The aim of this research is to propose a CORE (COmprehensive REasoning) framework for context-aware interactive system that aid to development of highly flexible, efficient, and interactive context-aware applications. Finally a prototype of the system will be implemented to validate the proposed framework.

Keywords - Context reasoning; Context-aware system; Context management; Context modeling; Quality of context

I. INTRODUCTION

A well-known definition of context is: “...any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the integration between a user and an application, including the user and the application themselves”. Context - Awareness (CA) is the ability of a computing device to sense, interpret, and interact with aspects of a user’s environment. A context-aware application has to be able to determine that the user is involved in different situations at different times. Situation - Awareness (SA) is considered as the particular kind of CA, where situation is viewed as logically aggregated pieces of context information. Current ubiquitous computing research tries to merge the material and digital worlds by incorporating physical and computing entities into smart spaces. It is hardly surprised that in near future every home, vehicle, and working place will be equipped with embedded and standalone computing devices such as sensors, augmented appliances, and actuators. This trend towards smart spaces is driving the need for applications that are interactive – able to interact with each other on behalf of user (to avoid disrupting user and let the user to concentrate on his task). For example, application in smart car can unobtrusively support driver, such as collecting destination information from his smart PDA, calculating shortest path to destination based on real time traffic information and his preference, updating current status of his elderly parent at home by communicating with smart home application. To perform such interactions in a seamless manner, the applications need to be aware of the situations which have significant impacts on the way they manage their adaptive behavior. A challenging task is inferring/deducing those situations/tasks (e.g., “Person’s sickness/heart-attack”) from lower-level context information (e.g., person’s pulse rate, body temperature, and so on) collected from environments equipped with error-prone devices. These situations may be associated with a certain level of uncertainty, depending on both the quality of the sensed information and precision of the deduction process. Moreover the interaction of the system not only depends on the derived situations but also the relationships among users (referred to as “social context”) and their preferences. For example, in smart office environments, during informal meeting with colleagues, calls coming from friends might be diverted to voice message but not those from parents. However, during formal meeting with executives and manager, all calls should be diverted to voice message.

Context-aware applications are designed to disrupt user as little as possible. The applications that know more about the user context are able to function efficiently and transparently adapt to the current user situation. It autonomously learns, automatically suggests what actions user prefers, and even sometimes acts on behalf of user, in designated situations. In case of wrong situation prediction, it may lead to disaster (e.g., vehicle system) or even death (e.g., elderly home care system). Hence, this is a challenge, since an application should interact intelligently and perfectly to everyday real life situations. However, there are several key issues to develop a context-aware system: (i) The well designed, generic and formal context modeling – to capture real world concepts and their relationships, facilitate context sharing, reusing and interoperability of applications; (ii) efficient and robust context management – to support context acquisition, storage, and query processing; and (iii) context reasoning – to resolve uncertainty in each process of context management, infer situations using context information facts and maps it to adaptive behavior.

The above discussion highlights the importance and challenges of context modeling, management and reasoning in context-aware systems. Particular interest is context
reasoning with uniform context modeling and robust context management to deal with intrinsic imperfection and uncertainty of context data, infer situations and map it to adaptive behavior to maintain a seamless interaction among systems in the environment.

II. MOTIVATIONAL SCENARIO

The proliferation of ubiquitous applications into smart spaces has fostered an increasing interest in context-aware applications. Here a scenario which comprises of several smart spaces (smart home, smart office, smart emergency service, and smart phone) is considered as shown in Fig. 1. It offers the prospect of significant improvements in the quality of life and level of care for elderly or disabled people.

![Figure 1. Pictorial representation of motivational scenario](image)

The scenario begins with Mary, the elderly mother of Jon, who had recently been hospitalized due to a heart attack. Although she had been discharged, the doctor recommends continuous monitoring of her condition. Upon return to her smart home with body sensors installed, her son, Jon will act as her principal caregiver. Assume that the context-aware application is already installed on Jon’s mobile/PDA and his home. The home application continuously communicates with the devices installed at home to predict Mary’s activities and sends alert message to Jon if any serious event happens. The application stores that information for future access, and provides an interface to query and access by other applications. The home application also has access to emergency services (e.g., ambulance, fire-service, police, etc).

Jon’s car also has been equipped with a context-aware application (called telematics). When he rides the car, the application collects destination information from his PDA’s schedule list based on priorities of the task. It communicates with road-side services and infrastructure, collects real-time traffic information, weather condition and road condition. Based on those information it calculates shortest path to destination. It also able to interact with Jon’s home application and keeps him updated with the real-time situations (what’s happening at home) while driving.

It happens that Mary falls down at home and gets so hard that she is unable to move and her pulse rate becomes abnormal, the home application detects that event and first tries to contact with Jon, but unfortunately Jon cannot respond as he is busy in a seminar, presenting in front of business executives and managers of the company. As a consequence the home application directly contacts with the emergency service and calls ambulance to bring Mary to hospital. It is important for the home application to have that functionality as current survey found that “more than two-thirds of Australians would not call an ambulance if they thought they were having a heart attack [1]”. The Ambulance picks Mary up to the hospital and informs the home application of the hospital address.

III. REQUIREMENTS OF CONTEXT-AWARE SYSTEMS

The context-aware interactive systems must address many of the requirements of traditional distributed systems, such as heterogeneity, mobility, and tolerance for component failures and disconnections. In addition, the system must fulfill the following requirements before it can become a reality:

- **Ease of deployment & configuration** – the system should be configurable to deploy it in different domains (e.g., home, office, vehicle, etc) to meet user and environmental requirements, potentially by non-experts.
- **Scalability & Interoperability** – It should be able to interact with other systems based on inferred situations, user preferences, and a set of user defined rules, norms and obligations. Its processing components and communication protocols must perform adequately for few to many sensors and actuators.
- **Efficient Context Management** – it should be able to store historical context information and have a query interface to access that information efficiently.
- **Robust Reasoning & interactive behavior** – it should be capable to derive situation from raw context information and adapt its behavior based on that.
- **Handling Uncertainty & QoC** – it should also have functionality to handle imperfect context information in each level of context processing.
- **Support security & privacy** – Flows of context information between the systems must be controlled according to user privacy needs and expectations.

Among all of the above requirements, context reasoning has great impact on applications interaction either through false positives or false negatives. For the above scenario, fail to detect Mary’s heart attack/sickness (false negative) might mean of her death or wrong situation prediction (false positive) may in turn cause financial loss (e.g., calling ambulance even in a good physical condition). It is a cardinal problem for ubiquitous applications. The aim of this research is to deal with dynamic, imprecise, inconsistence, contradictory, ambiguity, irrelevance, and incompleteness of context dimensions. The outcome of this research will be a framework for context-aware interactive systems, in conjunction with sophisticated context reasoning mechanism that will help application developer to develop a highly flexible, efficient, and interactive context-aware applications.

IV. LITERATURE REVIEW

There have been a number of context-aware systems available in the literature, from domain specific or location-based to general and extensible systems [2, 3]. For page
limitation, a review of most popular existing systems that span multiple layers of system architecture, especially discussed reasoning, has been shown in table I. It shows that comprehensive solutions do not exist. A further shortcoming is that existing approaches do not support system interaction.

**TABLE I. COMPARISON OF CONTEXT-AWARE SYSTEMS (KEY: + COMPREHENSIVE; ~ PARTIAL; - NONE)**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployment &amp; configuration</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Interoperability</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Context management</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Reasoning to infer high level context</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Reasoning</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Uncertainty &amp; QoC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In general, there are two main purposes for reasoning on uncertainty: improving the quality of context information, and inferring new kinds of context information. A number of mechanisms have been proposed in the literature for reasoning on uncertainty. Some of those approaches are: fuzzy logic, probabilistic logic, Bayesian networks, Hidden Markov models, and the Dempster-Shafer theory of evidence. Each of those approaches has its advantage and disadvantage. None is perfect. A comparative study of those approaches can be found in [9]. However, many problems have not yet been adequately addressed:

- How to model users’ relationships & preferences? And how to evaluate QoC and uncertainty?
- How to do more sophisticated reasoning to infer high level context/situation? And how to realize reasoning for social context?
- How to enable context-aware systems more interactive in presence of imperfect context?

**V. TOWARDS CORE FRAMEWORK**

This section presents the foundation of this work by summarizing the principles of the CORE framework.

**A. Conceptual Architecture**

An abstract layered architecture of CORE framework is shown in Fig. 2. The **Network layer** consists of protocol, sensors, devices or any context sources. The **Physical context management layer** responsible to acquire context data, transform it into the meaningful form for further inferring, and store it for later retrieval. The **Social context management layer** provides interface to define user relationships in terms of roles, constrains, and obligation. It also stores the context information and allows query processing. The **Adaptation layer** responsible for making decision, triggering action, and manage the user feedback/preferences. The main focus of this work will be layer 2 to 4 corresponding to above research questions.

![Figure 2. CORE architecture](image)

**B. Prototype Implementation**

Finally a prototype of the system based on the scenario will be implemented to validate the overall framework.

**VI. OUTLINE OF THE STUDY**

The problem formulation has done from Jul’09 to Jan’10. Existing techniques will be analyzed and a novel generic context modeling and evaluation technique will be proposed before Nov’10. Apart of that, context reasoning and management techniques will be investigated and a robust approach will be proposed by Oct’11. After that, a prototype will be implemented to validate the proposed approach by the end of Jan’12. For the final six months any “loose ends” on the research will be undertaken and the majority of the time will be spent for writing up thesis.

**ACKNOWLEDGMENT**

I would like to thank Prof. Jun Han (supervisor) and Dr. Alan Colman (co-supervisor) for their helpful guidance. This research is supported by the AutoCRC.

**REFERENCES**


An Incremental Methodology for Quantitative Software Architecture Evaluation with Probabilistic Models

Indika Meedeniya
Faculty of ICT, Center of Complex System and Services (CS3)
Swinburne University of Technology, Hawthorn, VIC 3122, Australia
Email: imeedeniya@swin.edu.au

Abstract—Probabilistic models are crucial in the quantification of non-functional attributes in safety-and mission-critical software systems. These models are often re-evaluated in assessing the design decisions. Evaluation of such models is computationally expensive and exhibits exponential complexity with the problem size. This research aims at constructing an incremental quality evaluation framework and delta evaluation scheme to address this issue. The proposed technique will provide a computational advantage for the probabilistic quality evaluations enabling their use in automated design space exploration by architecture optimization algorithms. The expected research outcomes are to be validated with a range of realistic architectures and case studies from automotive industry.

Keywords—Architecture evaluation, Probabilistic properties, Incremental Evaluation Models, Delta Evaluation

I. PROBLEM STATEMENT

Software intensive safety-and mission-critical systems are used in domains such as automotive/avionics, medical applications, railway etc. These systems have severe safety, reliability, performance and other non-functional requirements. Importantly, many of these attributes are probabilistic in nature. Probabilistic models are used in specification, evaluation and validation of such properties in system design.

Architecture optimization strategies have been proposed for finding better design alternatives with respect to the aforementioned quality attributes. Most of the optimization algorithms iteratively construct architectural variants with small changes by using strategies like re-deployment, redundancy allocation, component selection, and respective quality evaluation models are generated and re-evaluated for each variant. In this process, major limiting factors with probabilistic quality evaluation models are, i) computationally expensiveness in evaluation and, ii) exponential growth of complexity with respect to the size of problem [1]. Consequently, the use of probabilistic models in architecture evaluation has become a persistent challenge despite growing availability of computational resources as per Moore’s Law. For example in the automotive industry, re-evaluation of deployment architecture alternatives for safety, reliability and cost is a significant issue, resulting a major funding source for this project.

The proposed PhD research is aimed at addressing the problem by means of incremental construction of quality evaluation models and $\Delta$ Evaluation techniques for the model evaluation. Simply, given a change in adopted system’s architecture, the problem is to efficiently construct the evaluation model and evaluate the change in the probabilistic quality evaluation domain using the results of previous evaluations. Instead of re-constructing the entire evaluation model, it is proposed to propagate the change in the architecture to the probabilistic quality evaluation model. The re-computation is to be achieved by applying relevant $\Delta$ operators to the previous results.

II. RELATED WORK AND GAPS

A considerable amount of research has been performed in the area of software architecture evaluation with probabilistic models for properties such as Reliability [2], Safety [3] and Performance [4]. However, the evaluation models in the current approaches are constructed using concrete and problem specific relationships to the architectural model. They require complete reconstruction and re-evaluation of entire evaluation model even for a slight change in the architecture.

A stream of ongoing research can be seen in relation with evaluation models based on probabilistic transition systems, as the probabilities associated with the states and paths represent critically important aspects of the system. Being a subset of the above, parametric evaluation of Markov chains has gained significant interest in the research community, especially during the last few years. The recent work of Hahn et al. [1] has introduced a rational function based parametric evaluation approach which can be used in Discrete Time Markov Chains and Markov Reward Models. A predetermined set of transition probabilities in the Markov chain is allowed to be parametric in this approach and, reachability of specific state is calculated accordingly. In parallel with the above in the discrete time domain, Han et al. [5] have carried out similar work in the domain of real-time models, for Continuous Time Markov Chain (CTMC)s. They have formulated the computation of the probability of reaching a specific state in a CTMC, when the rates of the model are parametric. In summary, the current approaches are capable of evaluating Markov models when a predefined set of transition probabilities of the model is variable.
Apart from the limited coverage of evaluation models to date, a common limitation with current approaches is that they assume the model is fixed even though some transition probabilities may be parameters. However, in the process of evaluating alternative architectures this is not always the case. For example in a reliability model [2], the Markov model may also be altered when the architecture is slightly modified as new states may be added during the alternation. Thus, reconfiguration and re-evaluation of the Markov models with respect a change require further research.

III. PROPOSED SOLUTION AND CURRENT PROGRESS

The novelty of the proposed research is to enhance the evaluation of probabilistic quality models by enabling the propagation of a change from the software architecture to the quality evaluation domain. The overall approach can be viewed in two stages. Firstly, the static relationship between the architecture and probabilistic evaluation model will be changed by merging with the notion of Incremental model construction. When a change is made to the architecture, it will be propagated the evaluation model without a complete reconstruction. Secondly, the model evaluation process will be enhanced by applying the change through the model instead of complete model evaluation, entitled as Δ Evaluation. The research requires extensive analysis of commonalities in architecture evaluation models, their mathematical representation and relation to the architectural parameters. Selection of a feasible set of models would lead to the next phase of research, where proposed incremental evaluation frameworks will be applied. With the annotations in Figure 1 and using ⇒ to represent derivation, the approach can be characterized as:

\[(A, \Delta A, M) \Rightarrow \Delta M \Rightarrow M'\]

and

\[(M, \Delta M, Q) \Rightarrow Q'\]

As the initial step of applying incremental modeling, a set of distinct patterns has been identified in the probabilistic quality evaluation models: 1) Additive Models, 2) Directed Acyclic Graph(DAG)s and 3) Probabilistic Transition System(PTS)s. Challenges in applying incremental techniques are different in each category. For example in an additive NHPP(Non-Homogeneous Poisson Process) reliability model [2], the

\[\lambda_S(t) = \lambda_1(t) + \lambda_2(t) + \ldots + \lambda_n(t)\]

can be easily converted in to an incremental format,

\[\lambda_S^\prime(t) = \lambda_1^\prime(t) + \Delta(\lambda_i(t))\]

In comparison, for other models like DAG structured evaluation models, the applicability of incremental techniques is still a challenge. The relation of architectural parameters to respective mathematical formulation in DAGs and PTSs are being analyzed as of the current state of this research.

The potential contribution of this research will be a methodology to incrementally construct and evaluate probabilistic quality evaluation models. It is expected to have a significant benefit on the time needed to evaluate alternatives in architecture optimization. The proposed technique will also be applicable in Sensitivity Analysis, which is widely used as a technique to identify the impact of parameters to the composite system’s behavior. Parameter sweeps are most commonly applied in this purpose and, quality evaluation is carried out for each step of the sweep requiring re-evaluations. Runtime Adaptation is an emerging trend in software systems, where the inaccurate assumptions in the modeling phase are adjusted with the use of Runtime Monitoring. In this purpose, Run-time evaluation of quality metrics is required despite limited computational resources and time for the computation in real-time systems. This research foresees potential benefit in addressing computational challenges in the above.

IV. APPROACH FOR VALIDATION

An architecture evaluation and optimization framework, ArcheOpterix [6] has been implemented as an experimental test-bed for this research. The tool is currently capable of extracting embedded system specifications, evaluating probabilistic models and optimizing the architecture with respect to quality metrics using stochastic algorithms.

Experimental and theoretical consolidations are expected to be used in the validation of prospective research outcomes. A wide range of realistically generated architectures will be evaluated with existing models and incremental technique followed by comparison of execution time. Apart from the validation of computational gain, the accuracy of the incremental evaluation will also be affirmed. It is aimed to formulate a computational complexity comparison method for theoretical validation of the proposed incremental technique. Further, a set of challenging case studies from automotive software industry is expected to be included in

Figure 1. Outline of the proposed approach
order to verify that the proposed research addresses existing issues in real-world applications.

V. ACKNOWLEDGMENTS

This work was proudly supported by the Commonwealth of Australia, through the Cooperative Research Center for Advanced Automotive Technology (project C4-501: Safe and Reliable Integration and Deployment of for Automotive Software Systems).

REFERENCES


A Feature-Oriented Approach for Web Service Customization

Tuan Nguyen
Centre for Complex Software Systems & Services
Swinburne University of Technology, Melbourne, Australia
E-mail: tmnguyen@swin.edu.au

Abstract—This paper presents our ongoing work of a feature-oriented approach for Web services customization that helps to reduce the complexity in the customization process and enables the automated validation of customized services. Specifically, it exploits the feature modeling techniques of Software Product Line (SPL) to capture all possible customization options into a feature model. Weaving models, describing the links between the feature model and service models, enable service consumers to customize services through the selection of features in the feature model. By representing the capability of Web services at higher level of abstraction, i.e. feature level, it helps to simplify the customization process so that service consumers can focus on what features they want, rather than the technical details (e.g. operations, messages) of how to achieve it. In addition, automated analysis of feature models enables the automated validation of customized services.

Keywords: Service customization, feature model, variability modeling.

I. INTRODUCTION

An increasing number of organizations are turning to Service-Oriented Architectures (SOAs), based on Web service technologies, to consolidate and repurpose legacy applications, and combine them with new applications. Along with the proliferation of Web services, service ecosystems[1] are emerging, in which service providers interconnect their offerings in (un)foreseen ways to provide customized, value-added services. Services in Web service ecosystems have to satisfy various consumer demands. It is not likely that all consumers of a specific service have the same set of requirements on the service offered. Rather, such requirements are slightly different from one consumer to another. Therefore, supporting service customization is a crucial requirement in service ecosystems.

Service customization refers to operations of adapting a service to a particular application scenario. This is a non-trivial task that requires both technical knowledge and business expertise [2]. The challenges for an efficient service customization framework come from two perspectives. Firstly, actual business services support a plethora of possible customization options with a massive numbers of dependencies scattered between those options. The customization framework needs to efficiently address the complexity of the customization process. Secondly, due to previously explained complexity, service customization is a very error-prone process. The customization framework has to efficiently support the validation of customized services.

Software Product Line (SPL)[3] is a software engineering paradigm aiming at developing a family of software systems (or products) from reusable core assets. The key success of SPL is the use of feature models to model and manage variability in the product family. We argue that an efficient variability modeling technique is sufficient to address the challenges of service customization. In this paper, we propose an approach exploiting the feature modeling techniques of SPL to support service customization. Comparing to related works, our approach is more advantageous in term of supporting both reduced complexity and automated validation.

The paper is structured as follows. Section II describes our customization framework, its advantages, as well as the detail of how to realize it. We present related works in section III. We then conclude the paper and outline the future works in section IV.

II. SERVICE CUSTOMIZATION FRAMEWORK

A. Feature modeling technique and its benefits

In Software Product Line (SPL), a feature is defined as a visible characteristic of the product family [4]. A feature model represents the information of all possible products of a software product line in terms of features and relationships among them. A feature model is represented as a hierarchically arranged set of features composed by:

1. Relationships between a parent feature and its child features. Possible relationships are mandatory, optional and alternative.

2. Cross–tree (or cross–hierarchy) constraints (a.k.a. feature dependencies) that are typically inclusion or exclusion statements of the form: if feature F is included, then features A and B must also be included (or excluded).

Figure 1 depicts a simplified feature model representing a family of Video on Demand (VoD) services. According to the model, all customized VoD services must support Streaming, Device and Payment features. Messaging is optional feature that can be selected by a particular consumer. Consumers can select between high resolution (i.e. HighRes) and low resolution (i.e. LowRes), PC-oriented service (i.e. PC) and mobile-oriented service (i.e. MobilePhone), as well as payment by credit card (i.e. CreditCard) or through mobile carrier (i.e. MobileCarrier). In addition, because of feature dependencies, if a consumer selects mobile-oriented service, he has to use low resolution and payment by mobile carrier.
Feature is an effective communication "medium" among different stakeholders. It is often the case that business analysts and engineers speak of service characteristics in terms of "features the service has and/or delivers." Therefore, it is very intuitive to enable service customization based on feature models. More importantly, feature models enable modeling variability at the feature level, which generally is higher level of abstraction than implementation details (e.g. variability in operations or messages of service interfaces). Therefore, feature models greatly reduce the number of variation points available for customization. Consequently, it helps to reduce the complexity of service customization.

Moreover, researches in feature modeling techniques have achieved great advance in term of techniques for automated analysis [5]. The automated analysis of feature models is about extracting information from feature models using automated mechanisms. Examples of such extraction include but not limited to: deciding whether a feature configuration (i.e. feature selection) is valid or not; counting the number of possible products; applying filter to derive a subset of possible products; etc. Such automated analysis will help to automatically validate a customization.

B. Overview

![Service Customization Process](image)

Figure 2. Service Customization Process

In this section, we describe the overall steps of our service customization framework (Figure 2). At the first step, a service provider develops a customizable service as a product line. That is, the customizable service is developed as a family of services such that each family member is a possible customized service for a particular application scenario. To this end, the service provider captures the commonalities and variabilities of the service family into a feature model. The feature model will be used as the requirement model for developing service interface and service implementation. The detail of how to develop service interface and service implementation based on feature models will be explained later.

At the second step, the feature model is published to service registries as a part of service descriptions so that interested consumers can search for it. This step is different from traditional SOA in the term that the service variability model, i.e. the feature model, rather than the service interface description is published to service registries. The reason is, a service interface that a particular consumer consumes will be the result of a runtime customization process and it is not finalized at the time service is published to registries. At the third step, a service consumer discovers the feature model of interested services that it can customize.

The customization process at the consumer’s side starts from step 4. In this step, the service consumer selects desired features from the feature model. Feature selection operations include resolving variation points by enabling/disabling optional features and selecting a particular feature from an alternative feature group. Tooling support will help to simplify this process such that it will automatically resolve feature dependencies to prevent invalid feature selections or provide graphical interface for easy customization operations. The result of this step is a particular feature configuration.

Feature selection information is exchanged between the service consumer and the service provider (i.e. step 5) so that the service provider knows exactly which features the consumer desires. Based on this information, the service provider generates a particular service interface as well as deploys a particular service instance binding with that service interface (i.e. step 6). As the result of this, a particular WSDL description is returned to the consumer. At step 7, the service consumer will develop their own services integrating the customized service provided by the service provider.

C. Feature-oriented service development

In this section, we describe a model-driven approach to develop a customizable service conforming to a feature model (i.e. step 1 of our framework). The service engineering process is illustrated in Figure 3. Firstly, service developers model service capabilities to generate a feature model. Service capabilities in all possible application scenarios are structured into the feature model in which variation points represent customization options. Then, the service model template and process model template are developed using a superimposed variants approach of Software Product Line [6]. It is based on the idea of creating a model template containing all family members in a superimposed way. The specialization of such a template gives rise to different members (i.e. customized services or template instance) and it is carried out by purging model
template following the selection of features from a feature model, i.e. a feature configuration. Model template is itself a model expressed in the same notation as the template instance. Therefore, in our framework, service model template will be described using a service modeling language such as SoaML [7]. Likewise, process model template will be described using a process modeling language such as BPEL.

The superimposed variant technique requires mechanisms to relate features from the feature model to model template. In our approach, we use weaving models [8] for this purpose. Weaving models describe links between the feature model and other two models, i.e. service model template and process model template. Such weaving models help to capture relationships between model elements which will be used to generate a particular customized service.

### III. RELATED WORKS

Most existing approaches for service customization is limited to low-level technical aspects, such as the configuration of technical parameters for invocation and runtime. For such approaches, reducing complexity and automated validation remains grand challenges. [9] presents a policy-driven approach for service customization. However, this approach works with mostly informal policy descriptions that do not allow an automated validation of customization decisions. [2] also exploits variability modeling techniques from SPL to support service customization. Although the approach in [2] enables automated validation of customized services, variability is still modeled at low level of abstraction, i.e. addressing variability in operations, messages of service interface. Therefore, the approach can not address the complexity issue. Our approach is the only one that can address both challenges of service customization. Moreover, the approach also articulates the development process for service implementation conforming to the customizable service interface. This feature is also lack in both [2] and [9].

### IV. CONCLUSIONS AND FUTURE WORK

In this paper, we describe a feature-oriented approach for web service customization. Feature modeling techniques of SPL is exploited to address two grand challenges of service customization, i.e. reducing complexity and automated validation. The key concept is to use a feature model as the basis of service customization and utilizing weaving models to reflect customization decisions, i.e. feature configuration, to service interface and service implementation. We are going to develop a prototypical system to evaluate the feasibility of our approach.

### REFERENCES

Key Research Issues in Scientific Workflow Temporal Verification

Xiao Liu
Faculty of Information and Communication Technologies, Swinburne University of Technology
Hawthorn, Melbourne, Australia 3122
xliu@swin.edu.au

Abstract

In scientific workflow systems, temporal consistency is critical to ensure the in-time completion of workflow instances. To monitor and control the correctness of temporal consistency, temporal verification normally serves as one of the fundamental system functionalities. In recent years, great efforts have been dedicated to this area and it is high time that we should define its key research issues in order to keep our research on the right track. In this paper, we systematically investigate this problem and propose four key research issues based on the introduction of a novel temporal verification framework where representative solutions and current challenges are presented and discussed.

1. Introduction

Scientific workflow is a new special type of workflow that often underlies many large-scale complex e-science applications such as climate modelling, structural biology and chemistry, medical surgery or disaster recovery simulation [5]. Real world scientific as well as business processes normally stay in a temporal context and are often time constrained to achieve in-time fulfilment of certain scientific or business targets. Furthermore, scientific workflows are usually deployed on the high performance computing infrastructures, e.g. cluster, peer-to-peer and grid computing, to deal with huge number of data intensive and computation intensive activities. Therefore, as an important dimension of workflow QoS (Quality of Service) constraints, temporal constraints are often set at build time and verified at run time to ensure in-time completion of scientific workflow executions. Temporal verification, as one of the fundamental workflow system functionalities, is often implemented to monitor run-time temporal consistency state in order to maintain temporal correctness and handle temporal violations proactively before heavy-weight exception handling needs to be triggered.

In recent year, temporal verification becomes an important research topic and attracts many efforts from scientific workflow research area due to its distinctive differences from traditional business workflows. Currently, most work focuses on the monitoring of large scale scientific workflows. However, as will be discussed in Section 3, temporal verification is not an independent task. To guarantee temporal correctness of scientific workflow execution, four important consecutive tasks including constraint setting, checkpoint selection, temporal verification and temporal adjustment should be implemented. Therefore, in this paper, to keep our research on the right track and focus more on the open issues and challenges, a novel temporal verification framework is proposed to investigate the key research issues.

The remainder of the paper is organised as follows. Section 2 presents a motivating example. With a novel temporal verification framework, Section 3 proposes four key research issues with their representative solutions and current challenges. Finally, Section 4 addresses the conclusion.

2. Motivating Example

The entire weather forecast workflow contains hundreds of data intensive and computation intensive activities. Major data intensive activities include the collection of meteorological information, e.g. surface data, atmospheric humidity, temperature, cloud area and wind speed from satellites, radars and ground observatories at distributed geographic locations. These data files are transferred via various kinds of network. Computation intensive activities mainly consist of solving complex meteorological equations, e.g. meteorological dynamics equations, thermodynamic equations, pressure equations, turbulent kinetic energy equations and so forth which require high performance computing resources. Due to the space limit, it is not possible to present the whole forecasting process in detail. Here, we only focus on one of its segments for radar data collection. As depicted in Fig. 1, this workflow segment contains 12 activities which are modeled by Stochastic Petri Nets. For simplicity, we denote these activities as $X_1$ to $X_{12}$.

![Figure 1. Example scientific workflow segment](image)

It is evident that the duration of these scientific workflow activities are highly dynamic in nature due to their data complexity and the computation environment. However, to ensure the weather forecast can be broadcast on time, every scientific workflow instance must be completed within specific time durations. Therefore, in the first place, temporal constraints must be set to control the overall workflow execution time. For our example workflow segment, to ensure that the radar data can be collected in time and transferred for further processing, at least one overall upper bound temporal constraint $U(X_1, X_{12})$ with the value of $\omega(X_1, X_{12})$ is required to be set at the place follows $X_{12}$. After that, during workflow run time, temporal verification should be conducted on some selected activity points to check the temporal correctness of given constraints. But the problem is
where, i.e. which activity point from $X_1$ to $X_{12}$ should be selected. Evidently, it is of low efficiency to check on every activity point. Furthermore, if we selected an activity point such as $X_5$, given the temporal constraint of $u(X_1, X_{12})$, how to describe the current temporal consistency state, namely the trend deviates from or approaches to temporal correctness, is also a problem. Finally, if on $X_5$, we verified a high probability of temporal violation for $u(X_1, X_{12})$, there is another problem concerning with how to handle the current inconsistency state so as to ensure the subsequent activities $X_5 - X_{12}$ could be finished within the remained time after the completion of $X_5$. Otherwise, the collection process for radar data will be delayed and further deteriorate the in-time completion of the entire weather forecasting process.

3. Key research issues

3.1. Temporal verification framework

As depicted in the outlier of Fig. 2, temporal verification framework consists of four consecutive tasks including constraint setting, checkpoint selection, temporal verification and temporal adjustment. The inner circles stands for three dominant factors concerned with temporal verification including temporal constraints specified in workflow models, the dynamic performance of underlying services and the system historic data of scientific workflow systems. Detailed discussions for each task are presented in the following sections.

![Figure 2. Temporal verification framework](image)

3.2. Constraint Setting

The first task of the temporal verification framework is constraint setting which specifies temporal constraints in workflow models at build time. Most current work holds the assumption that temporal constraints are pre-defined and focuses only on run-time temporal verification while neglects the fact that efforts put at run-time will be mostly in vain without build-time setting of high quality temporal constraints. The reason is obvious since the purpose of temporal verification is to identify potential violations of temporal constraints to minimise the exception handling cost. Therefore, if temporal constraints are of low quality themselves, temporal violations are highly expected no matter how much efforts have been dedicated by temporal verification. Generally, the quality of temporal constraints can be measured by at least two criteria: 1) well balanced between user requirements and system performance; 2) well supported for both overall coarse-grained control and local fine-grained control.

One of the representative solutions for setting temporal constraints in scientific workflow is a probabilistic strategy as proposed in [3]. With the probability based temporal consistency, the basic idea of this strategy is first obtain the normal distribution of the overall completion time through the weighted joint normal distribution of all activity durations. After that, a negotiation process is provided for the user and system manager to specify the coarse-grained constraint of the entire workflow and then a propagation process is implemented to specify the fine-grained constraints for each workflow activity. This strategy is very competitive in terms of both efficiency and effectiveness.

One of the current challenges in constraint setting is that the locations of temporal constraints are normally assumed to be predefined but actually unknown in the real world. It is evident that the locations of temporal constraints have great impact on the efficiency control of workflow executions. For example, a constraint set on the critical path is more reasonable and effective than the one set on sub paths. Therefore, where to set temporal constraints so as to maximise the effectiveness of temporal verification is an interesting research topic.

3.3. Checkpoint selection

The second task of the temporal verification framework is checkpoint selection which dynamically selects activity points along workflow execution to check the current temporal consistency state. Since it is of extremely low efficiency and high cost to conduct temporal verification on each activity point, checkpoint selection is a necessary pre-task. Conventional work selects special activity points such as the start activity, the end activity or the decision point as checkpoints. However, the basic criteria for checkpoint selection are necessity and sufficiency. Here, necessity means that only those activity points where real temporal inconsistency states take place are selected and sufficiency means that there are no any omitted points.

One of the representative solutions for checkpoint selection is a minimum time redundancy based checkpoint selection strategy which utilises multiple discrete states based temporal consistency. Take the verification of SC (strong consistency) constraints for example, the basic idea of this strategy is to find out the activity point $a_p$ where $R(a_p) > D(a_p) + MTRSC(a_{p-1})$. Here, $R(a_p)$ is the run-time duration, $D(a_p)$ is its maximum duration and $MTRSC(a_{p-1})$ is the minimum time redundancy at $a_{p-1}$ [2]. Then, $a_p$ is selected as a checkpoint and all the previous SC constraints need to be verified. This strategy is proved to be of both necessity and sufficiency.

One of the current challenges for checkpoint selection is its efficiency. Although the criteria of necessity and sufficiency have significantly reduced the cost over the previous strategies, it is still huge especially in a scientific workflow of thousands of activities. Therefore, how to minimise the cost while keeping satisfying effectiveness of temporal verification, is a challenging problem. Recently, a paper addresses this problem by investigating the dependencies of fixed-time constraints [2]. Specifically, with temporal dependency, the consistency state of some later constraints can be deduced from previous ones do not need to take
any checkpoints. Therefore, a number of checkpoints are omitted. However, many other aspects such as the activity dependency and workflow structure dependency can be further explored.

3.4. Temporal verification

The third task of the temporal verification framework is temporal verification which checks the current temporal consistency state according to the definition of temporal consistency. Evidently, the actual temporal verification task is very intuitive but has strong relationship with its previous tasks. Therefore, temporal verification is normally discussed together with the work on specific definitions of temporal consistency or checkpoint selection strategies.

One of the representative solutions introduced in [1] is for multiple states based temporal consistency where the relationships between the verification of SC, WC, WI and SI are discussed. The basic idea is that for a specific temporal constraint, if it is verified to be a higher consistency states, e.g. SC, then we do not need to check its WC, WI or SI, so on and so forth. This rule saves unnecessary cost on temporal verification.

The challenges of temporal verification mainly come from checkpoint selection since they are always performed together. Evidently, the efficiency of and effectiveness of checkpoint selection directly affects the performance of temporal verification. Therefore, as discussed in Section 3.3, how to improve the efficiency, i.e. saving the execution time and computation cost, is a challenge.

3.5. Temporal adjustment

The last task of the temporal verification framework is the temporal adjustment. Current work on checkpoint selection and temporal verification can only deal with the detection of temporal violations, but an important follow-up task is how to handle those violations, i.e. adjusting temporal inconsistency states. So far, the study on adjusting temporal inconsistency state in scientific workflows, or temporal adjustment for short, is very limited.

Temporal adjustment can be regarded as a kind of exception handling in workflow systems. Generally speaking, a process such as the one recruits additional resources is deemed as conventional exception handling while the one without is regarded as temporal adjustment. However, temporal violation is quite different from conventional exceptions in workflow systems which are mainly on fault tolerance of activity failures. [4] introduces five types of workflow exceptions where temporal violation can be classified into deadline expiry. Meanwhile, three alternate courses of recovery action, i.e. no action (NIL), rollback (RBK) and compensate (COM), are also presented. As for temporal violation, COM, i.e. time deficit compensation, is a suitable recovery action.

One of the representative solutions introduced in [1] is a time deficit allocation strategy (TDA) which compensates current time deficit by utilising the expected time redundancies of following activities. However, since time deficit has not been truly reduced, this strategy can only delay the violations of some local constraints, but has no effectiveness on overall constraints, e.g. the deadlines. On the contrary, workflow rescheduling can indeed make up time deficit by expediting the execution of those non-commenced workflow activities. However, since workflow scheduling is a NP complete problem, extra cost is hence inevitable [6].

Up to now, there are few efforts dedicated to this topic. However, to form a practical and effective temporal verification framework, temporal adjustment plays a significant role to actually correct the temporal incorrect states. Therefore, it is high time that a temporal adjustment strategy should be investigated. As for its measurement, the major criterion is its performance, i.e. how much time deficit can be eliminated. Meanwhile, the cost of the compensation process (including both monetary and overhead cost) should be considered since it is unreasonable if the cost of the compensation process itself exceeds the expected cost to be brought by temporal violations. However, how to avoid conventional exception handling with a cost-effective temporal adjustment strategy is a challenge.

4. Conclusion

Temporal verification is one of the most important topics in the research area of scientific workflows. To keep our research on the right track, this paper has presented four key research issues of scientific workflow temporal verification, i.e. constraint setting, checkpoint selection, temporal verification and temporal adjustment. Actually, these four research issues are four consecutive tasks defined in the novel temporal verification framework. In this paper, at least one representative solution and one current challenge for each research issue have been presented and discussed. Specifically, the major challenge for constraint setting is the location of temporal constraints, the primary challenge for both checkpoint selection and temporal verification is the problem of efficiency, and the main challenge for temporal adjustment is the measurement of cost-effectiveness.

To conclude, considering its current state and with all these challenges ahead, scientific workflow temporal verification is still in its infancy. More efforts are required to explore this area, especially on these four key research issues identified in this paper.

5. References


Integration Method for Policy Integration

Trong Hieu Tran
Centre for Complex Software Systems and Services (CS3)
Swinburne University of Technology
Melbourne, VIC, Australia
hhtran@groupwise.swin.edu.au

Abstract—Policy-based security is an effective approach to manage knowledge systems by handling all behaviours of a system through a set of rules. However, this approach also has to cope with potential conflicts in administrative processes. In this paper, we will propose a new approach to solve conflicts and to integrate rules in a policy. A new representation of rules is given, the distances between rules are defined as well as postulates are presented and analysed. Algorithms for integrating policy also have been proposed and examined.

Keywords: Policy-based security, knowledge integration

I. INTRODUCTION

Security is one of the important problems in multi-agent systems as well as database systems. It has become increasingly important for computer and information systems with explosive growth of the Internet and the widespread use of wireless networks. There are some methods to ensure the security for a system, and policy-based management is one of the most common and effective approaches. By this approach, we can set the configuration easily, have the flexibility in development and maintenance processes as well as define general tasks in a high level without knowing about the detailed specification of system in which policy is applied. This approach is applied in sub-fields of AI and database systems such as security and access management [9], network management and monitoring [6], and electronic commerce [2, 7].

In policy-based management approach, all behaviours of a system are handled by a sequence of rules called policy. This approach has some advantages. However, if an inconsistent situation arises, it may lead the system to an unknown state or an error. Unfortunately, this situation is difficult to avoid because the rules of a policy may be given by many administrators, in different periods of time, and without the clear idea of their purposes [1]. Therefore, working out a solution for security policy integration problem is one of the basic requirements of the system administrating.

In order to integrate policies, the most common task that we have to perform is solving conflicts. There are several methods for conflict resolution proposed. In [12], authors propose the method base on Consensus Theory to resolve conflicts and integrate security policy. The methods, based on the order of rules in the policy, the priority of the restriction of rules, and the most/least specific condition, are introduced in [11]. Some structures defined to represent policies have been also examined. In [1] the authors defined an algebra of security policy as well as its semantics to combine authorization specifications. The hierarchy structures are used in [3, 4] and graph is recommended in [5, 10]. The representation of security rules on the syntactic level has been surveyed and analysed in works published recently. For instance, set-based approach and semi-lattices are used to solve conflicts in policy rules [1] or relational structures have been proposed in [12]. In this paper, we propose the approach to represent this kind of knowledge and to solve conflicts on a logical semantic level.

The rest of this paper is structured as follows. In Section 2, we present some related concepts such as some definitions of rule, policy, and conflict. The distance functions between rules are introduced in Section 3. The postulates are proposed and some algorithms are examined in Section 4. At last, some conclusions are included in Section 5.

II. BASIC NOTIONS

Definition 1: A rule is tuple $R = \langle C, A \rangle$ where $C$ is a family of conditions, $A$ is a set of actions.

We concentrate about the representation of conditions of rules as follows:

The real world of conditions includes a set $C = \{c_1, c_2, ..., c_n\}$ of fields of conditions and a set $V = \{V_{c_1}, V_{c_2}, ..., V_{c_n}\}$ of the elementary values of condition fields respectively, (each $V_{c_i}$ is the set of values of condition filed $c_i$, or $V_{c_i}$ is super domain of $c_i$). Shortly, pair $(C, V)$ is called a real world of conditions. Let $\prod(V)$ denote the set of all subsets of set $V$. We also assume that for each condition field $c_i$ its value is always a set of elementary values from $V_{c_i}$ and obviously, it is an element of set $\prod(V)$. An elementary value means a value, which is indivisible in the system.

An expression $(c = v)$ or $(c \neq v)$ where $c \in C$, $v \in \prod(V)$ and $v$ is a finite set, is called a literal from real world $(C, V)$. If a literal has form $(c = v)$ we call it a positive literal if it has form $(c \neq v)$ then we call a negative literal. A negative literal $(c \neq v)$ can be considered to be equivalent to $\neg(c = v)$. A negative literal may be transformed into a positive literal by using the attribute super domains, that is literal $(c \neq v)$ is equivalent to literal $(c = v')$ where $v' = V_{c_i}\backslash c_i$. By $C_{CV}$ we denote the set of all conditions of $(C, V)$-based literals.

Definition 2. By the semantics of conditions, we define the following function:

$$S_{C_v} : C_{CV} \rightarrow \bigwedge_{x \in C_{CV}} V_x$$

such that $S_{C_v}(x) = \{(a_1, a_2, ..., a_n) : a_i \in v_i, i = 1, n\}$

where $x = (c_1, v_1) \land (c_2, v_2) \land ... \land (c_n, v_n), v_i \in \prod(V_i)$
Thus the semantics of condition \( x \) is the set of all tuples built by Cartesian product of all super domains of the condition fields occurred in \( x \). The intuition of this definition is based on the aspect that if condition \( x \) represents the condition of a rule, set \( S_C(x) \) will consist of all possible scenarios which are included in \( x \).

**Definition 3.** The semantics of rules is the semantics of conditions binding with the corresponding actions.

Because of the assumption that each rule has only one action, it is intuitive to consider that the semantics of a rule includes all possible scenarios of the condition binding with the action. The following example illustrates the intuition:

**Example 1.** Considering to a simple rule in access filter policy of a system as follows:

\[
 r = ((\text{protocol}, \{TCP\}) \land (\text{IP_address}, \{192.168.0.2-4\}) \land (\text{port}, \{100, 102\}), (\text{action} = \text{Permitted})).
\]

The semantics of rule \( r \) has the following tuples:

<table>
<thead>
<tr>
<th>Protocol</th>
<th>IP_address</th>
<th>Port</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>192.168.0.2</td>
<td>100</td>
<td>Permitted</td>
</tr>
<tr>
<td>TCP</td>
<td>192.168.0.2</td>
<td>102</td>
<td>Permitted</td>
</tr>
<tr>
<td>TCP</td>
<td>192.168.0.3</td>
<td>100</td>
<td>Permitted</td>
</tr>
<tr>
<td>TCP</td>
<td>192.168.0.3</td>
<td>102</td>
<td>Permitted</td>
</tr>
<tr>
<td>TCP</td>
<td>192.168.0.4</td>
<td>100</td>
<td>Permitted</td>
</tr>
<tr>
<td>TCP</td>
<td>192.168.0.4</td>
<td>102</td>
<td>Permitted</td>
</tr>
</tbody>
</table>

We have the following properties of the semantics of conditions:

**Proposition 1.** Conditions

\[
 x = (c_1, v_1) \land (c_2, v_2) \land \ldots \land (c_n, v_n) \text{ and } x' = (c_1, v_1) \land (c_2, v_2) \land \ldots \land (c_n, v_n) \land (c, v_c)
\]

where attribute \( c \) does not occur in \( x \), should have the same semantics, that is \( S_C(x) = S_C(x') \).

Conditions \( x \) and \( x' \) having the same semantics are called equivalent to each other.

**Definition 5.** Rules \( r_1 = (c_1, a_1) \) and \( r_2 = (c_2, a_2) \) are conflict if \( S_C(c_1) \cap S_C(c_2) \neq \emptyset \).

The conflict between two rules occurs in the case if it there exists scenarios in which the rules have the same condition. In work [12], authors classified and analysed types of policy conflicts based on the relations between rules such as shadowing conflict, redundancy conflict, correlation conflict, and exception conflict.

**Definition 6: A policy is a sequence of rules, which is used to administer, manage, and control access to a security system [1].**

Formally, we denote a policy \( p \) including rules \( r_1, r_2, \ldots, r_n \), and \( r_n \) by a sequence as follows:

\[
p = < r_1, r_2, \ldots, r_n >
\]

By symbol "\( \oplus \)" we denote the concatenation between two policies, so a policy can be built as follows:

1. If \( r \) is a rule, \( < r > \) is a policy.
2. If \( p_1 \) and \( p_2 \) are policies, \( p_1 \oplus p_2 \) is also a policy.

It is easy to notice that with \( p_1, p_2 \) and \( p_3 \) are policies; the concatenating operator \( \oplus \) has following characteristics:

a) \( p_1 \oplus p_2 \neq p_2 \oplus p_1 \)

b) \( p_1 \oplus p_2 \oplus p_3 = p_1 \oplus (p_2 \oplus p_3) \)

**III. DISTANCES BETWEEN SECURITY RULES**

Generally, the distance between two rules may be understood as the sum of the distance between the conditions and the distance between the actions of these rules. It is intuitive that the distance between two conditions should be equal the minimal cost of translating the semantics of the first condition into the semantics of the second one. Thus we have:

**Definition 7.** For conditions \( b = c_1 \land c_2 \land \ldots \land c_v \) and \( b' = c_1' \land c_2' \land \ldots \land c_v' \), their distance \( d_C(b,b') \) is equal the minimal cost for transforming set \( S_C(b) \) into set \( S_C(b') \).

By the operation transforming set \( S_C(x) \) into set \( S_C(x') \) we mean performing such operations as adding, removing and transformation to the elements of set \( S_C(x) \), which in the result give set \( S_C(x') \). For the need of the definition of these operations, we define the following cost functions:

- Function \( d_I: V \rightarrow [0, +\infty) \): specifies the cost for adding (or removing) of an elementary value to (or from) a set.
- Function \( d_C: V \times V \rightarrow [0, +\infty) \): specifies the cost for transformation of one elementary value into another.

Similarly, like in work [8] for functions \( d_I \) and \( d_C \) we also accept the following assumptions:

a) Function \( d_I \) is a metric, i.e. for any \( x, y, z \in V \) the following conditions are held:
   - \( d_I(x, y) \geq 0, d_I(x, y) = 0 \) if and only if \( x = y \),
   - \( d_I(x, y) = d_I(y, x) \),
   - \( d_I(x, y) + d_I(y, z) \geq d_I(x, z) \).

b) For any \( x, y \in V \) \( d_C(x, y) \leq d_I(x, y) \leq d_I(x) + d_I(y) \).

For convenience in calculating, in this work we assume that \( d_I(x) = d_I(y) = 1 \) and \( d_I(x, y) = d_I(x) + d_I(y) \).

**Definition 8.** For rules \( r_1 = (c_1, a_1) \) and \( r_2 = (c_2, a_2) \), the distances between \( r_1 \) and \( r_2 \) is calculated as

\[
d_C(r_1, r_2) = d_C(c_1, c_2) + d_C(a_1, a_2)
\]

where \( d_C(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \) (\( x, y \in A \)).

**IV. POSTULATES AND ALGORITHMS**

Let \( U \) be a finite universe consisting of rules may occur in a policy system. By \( \Pi(U) \) we denote the set of all finite and nonempty sequences with repetitions of set \( U \). Each element of \( \Pi(U) \) is called a conflict profile in policy system. Therefore, a conflict profile is a set with repetition of rules with a determined order, in other words, it is called a policy profile in some system. By integration function, we mean the following function:

\[
 C: \Pi(U) \rightarrow 2^U
\]

In this function, we assume that the result will be a sequence without repetition. For a profile \( X \) sequence \( C(X) \) is called the integration of \( X \). By \( C(U) \) we denote the set of all integration functions for universe \( U \).

**Definition 9.** By an integration function \( C \in C(U) \) for profiles of rules we understand a function:

\[
 C: \Pi(U) \rightarrow 2^U
\]
which satisfies one or more of the following postulates:

P1. For \( X = \{x_1, x_2, \ldots, x_n\} \), \( x_i \in U \), there should be:
\[
\mathcal{C}(X \oplus \Delta \ominus) = \mathcal{C}(X)
\]

P2. For \( X = X_1 \oplus X_2 \) and \( Y \) is a subsequence of \( \mathcal{C}(X) \), there should be
\[
\mathcal{C}(X) = \mathcal{C}(X_1 \oplus Y \ominus X_2)
\]

P3. \( \mathcal{C}(X) \neq \varnothing \) for any profiles \( X \)

P4. For \( X = X_1 \oplus X_2 \), there should be
\[
\mathcal{C}(X) = \mathcal{C}(C(X_1) \oplus \mathcal{C}(X_2))
\]

P5. A consensus \( x^* \in \mathcal{C}(X) \) should minimize the sum of distances (\( O_1 \)-consensus):
\[
\sum_{x \in X} d(x^*, x) = \min_{x \in X} \sum_{x \in X} d(x^*, x);
\]

P6. A consensus \( x^* \in \mathcal{C}(X) \) should minimize the sum of distances (\( O_2 \)-consensus):
\[
\sum_{x \in X} d^2(x^*, x) = \min_{x \in X} \sum_{x \in X} d^2(x^*, x)
\]

Some commentary of these postulates is given as follows:

- Postulate P1 implies that we can remove repetition subsequences of rules.
- Postulate P2 states that the integration sequence of the profile \( X \) is robust.
- Postulate P3 implies that there is always solution for any integration process.
- The idea of postulate P4 is based on divide-and-conquer strategy, a very common one in Artificial Intelligent.
- Postulates P5 and P6 refer to the popular criteria of consensus theory used to determine the integration result quantitatively.

With the assumption that all condition fields are independent and all actions are independent, we present an algorithm to integrate policy rules based on O1 criterion as follows:

**Algorithm 1:** Computing \( O_1 \)-consensus \( X \) for policy \( P \).

**Given:** A policy \( P = \{x_1, \ldots, x_n\} \) of \( (C, A) \)-based rules.

**Result:** Consensus \( X \) for \( P \) satisfies P1, P2, and P5.

**BEGIN**

1. Create the set: \( PS = \bigcup_{y \in P} S(y) \) where \( S(x) \) is the set of semantics of rule \( x \);
2. Let \( X := \varnothing; S_p := \infty; Z := PS \);
3. While \( Z \neq \varnothing \) do
   1. Select from \( Z \) an element \( z \) such that the sum
      \[
      \min_{min\_value} := \sum_{y \in P} \min_{z \in (y)} d(x, y)
      \]
      is minimal;
      \( Z := Z \setminus \{z\} \);
   2. If \( S_p \geq min\_value \) then
      Begin
      \( S_p := min\_value; \)
      \( X := X \cup \{z\}; \)
      End;
   End While;
4. Return \( X \);
**END**

The idea of this algorithm is stated as follows: we firstly collect all the semantics of all rules in set \( P \) of rules, and then we step by step choose the best semantics, of which the distance among it and chosen ones is minimum and the reached value is smaller than current total distance value, and add it in chosen set of semantics. We perform these steps until all semantics in initial set is examined.

The computational complexity of Algorithm 1 is evaluated based on the loop \( \text{While} \) in step 3 and the evaluation of \( min\_value \) in step 3.1. Thus, the computational complexity of \( O(n^3) \) where \( n = \text{card}(PS) \).

**V. CONCLUSIONS**

In this paper, we have proposed a new approach to solve conflicts and to integrate rules in a policy, in which a new representation of policy rules on the semantic level has been presented and the distances between rules are defined. Several postulates for policy integration are proposed and analysed. An algorithm for policy integration has been proposed and examined. In future works, we will continue to have more deeply analysis about this approach and work out some algorithms satisfying other groups of chosen postulates.

**ACKNOWLEDGMENT**

The author would like to thank Professor Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland and Professor Ryszard Kowalczyk and Doctor Bao Vo, Swinburne University of Technology, Australia for the instruction and supervision to complete this paper.

**REFERENCES**


Selectivity Estimation for SPARQL Graph Pattern

Hai Huang
Faculty of ICT
Swinburne University of Technology
Melbourne, Australia
hhuang@swin.edu.au

Abstract—The Resource Description Framework (RDF) is a standard format for encoding machine-readable information on the Semantic Web. Recently, more and more data is being stored in RDF format. A fundamental problem related to RDF query processing is selectivity estimation, which is crucial to RDF query optimization for determining a join order of RDF triple patterns. In this paper, we focus research on selectivity estimation for SPARQL graph patterns.

The previous work takes the join uniformity assumption when estimating the joined triple patterns. This assumption would lead to high inaccurate estimation in the cases where properties in SPARQL graph patterns are correlated. In our research, we take into account the dependencies among properties in SPARQL graph patterns and propose a more accurate estimation model. Due to the fine-grained characteristic of RDF modeling, star and chain query patterns are common in SPARQL graph patterns. Thus, we first focus on these two basic patterns and propose to use Bayesian network and chain histogram respectively for estimating the selectivity of them. Then, for estimating the selectivity of an arbitrary SPARQL graph pattern, we design algorithms for maximally combining the results of the star and chain patterns we have precomputed.

Keywords—Selectivity Estimation, RDF Query Processing, Bayesian Network

I. INTRODUCTION

The Resource Description Framework (RDF) is a standard format for encoding machine-readable information on the Semantic Web. Recently, more and more data is being stored in RDF format. RDF data is a set of triples and each triple called statement is of the form (subject, property, object). RDF data can also be represented as a graph with nodes representing resources or their property values and labeled arcs representing properties. This data representation is general and flexible. Almost any kind of data can be represented in this format. However, this fine-grained model leads to queries on RDF data with a large number of joins, which is an inherent characteristic of querying RDF data [1].

Since the use of RDF to represent data has grown dramatically over the last few years, query processing on RDF data becomes an important issue in realizing the semantic web vision. Some query languages such as SPARQL [2] have been developed. As we know, accurate estimation of the result size of queries is crucial to query processing. Cost-based query optimizers use estimated intermediate result size to choose the optimal query execution plan. In addition, selectivity estimation can also be used to approximately answer counting (aggregation) queries and to allow efficient load balancing for parallel join on multiprocessor systems.

As a SPARQL query has a large number of joins, estimating precisely the joined triple pattern is very important. Some work has been done in this area. In [5], [6] the join uniformity assumption is made when estimating the joined triple patterns with bound subjects or objects (i.e., the subjects or objects are concrete values). They assume that each triple satisfying a triple pattern is equally likely to join with the triples satisfying the other triple pattern. However, this assumption does not fit reality in many cases. And when the data are inconsistent with this assumption, the resulting selectivity estimation could differ from the true selectivity by orders of magnitude.

For example, in Figure 1, a SPARQL query is posed on an RDF database, which retrieves academic staff members and the courses they teach with some conditions. Suppose we want to estimate the selectivity \( sel(t_1 \times t_3) \) of two joined triple patterns \( t_1 \) and \( t_3 \). It would overestimate the result size \( sel(t_1 \times t_3) \) using the formula (1) proposed in [5] as follows:

\[
sel(t_1 \times t_3) = \frac{S_p}{|T|^2} \times sel(\text{Income}, \ '\leq 70K') \times sel(\text{Position}, \ 'Prof.')
\]

(1)

Figure 1. RDF SPARQL query
where $S_P$ is the result upper bound of the joined triple patterns $(\text{TZ}, \text{Income}, \text{?W})(\text{TZ}, \text{Position}, \text{?Y})$ and $|T|$ is the number of triples in the database. $sel(\text{Income},'\leq 70K'$ and $sel(\text{Position},'\text{Prof}.$) are the object selectivity of $t_1$ and $t_3$. It assumes that each triple satisfying pattern $t_1$ is equally likely to join with triples satisfying triple pattern $t_3$. But in fact, the triple matching $t_3$ which indicates that the person is a professor who is supposed to have higher income (‘$> 70K$’). Thus, the triples matching $t_1$ are less likely to join with triples matching $t_3$.

There are two observations from this example. First, we can observe that the join uniformity assumption is not applicable in those cases where properties have correlations (dependencies) with each other. In fact, this assumption is rarely satisfied in real data, so we need a more accurate model to relax this assumption for estimating the result size of a SPARQL query. Second, the SPARQL query in Figure 1 is a composition of a chain pattern and a star pattern, and these two types of patterns are very common in SPARQL queries. It is desirable to find appropriate ways to estimate the selectivity of these common patterns.

II. EXPECTED CONTRIBUTION OF THIS RESEARCH

We try to solve the problem of selectivity estimation for SPARQL graph patterns. The contributions of this research can be summarized as follows:

- We study the star and chain patterns with correlated properties and propose two methods for estimating the selectivity of them respectively. For star query patterns, we construct Bayesian networks to compactly represent the joint probability distribution over values of correlated properties. And for chain query patterns, we build the chain histogram, which can obtain a balance between the estimation accuracy and space cost.

- For an arbitrary SPARQL query represented as a graph pattern, we propose algorithms for maximally combining the results of chain patterns and star patterns that we have precomputed to estimate the overall selectivity of the graph pattern.

III. RELATED WORK

In [5], the authors propose the framework of static Basic Graph Pattern (BGP) optimization based on selectivity estimation. They devise a number of heuristics for the selectivity estimation of joined triple patterns. The heuristics range from simple variable counting techniques to more sophisticated selectivity estimations based on the probabilistic framework. In [6] the authors propose two kinds of statistics for selectivity estimation. The first one, specialized histograms, is generic and can handle any kind of triple patterns and joins with independence assumption. The second statistics computes frequent join paths in the data, which are similar to our methods. Both [5], [6] take the join uniformity assumption when estimating the selectivity of joined triple patterns. We avoid this assumption and propose a more accurate estimation model that takes correlations among properties into account.

IV. METHODOLOGY

A. Estimation for star patterns using Bayesian networks

The star graph pattern is common in SPARQL graph patterns. It has the form of a number of triple patterns with different properties sharing the same subject (an example is shown in Figure 2). For estimating the selectivity of frequent star patterns, we construct the cluster-property table $R$ for each one.

![Figure 2. Star-style graph pattern](image)

Given a frequent star pattern $Q$ with predicates $prop_1$, $prop_2$, $\ldots$, $prop_n$, if we know the joint probability distribution over values of properties $Pr(prop_1 = o_1, prop_2 = o_2, \ldots, prop_n = o_n)$ in $R$, we can easily obtain the selectivity $sel(Q)$ of $Q$ as: $sel(Q) = Pr(prop_1 = o_1, prop_2 = o_2, \ldots, prop_n = o_n) \cdot |R|$, where $|R|$ is the number of rows in $R$. However, it is impossible to explicitly store the joint probability distribution over property values in $R$, since the possible combinations of values of properties could be exponential. Thus, we need an appropriate structure to approximately store the joint probability distribution information. We employ Bayesian network [3] to store the joint probability distribution information. Bayesian networks make use of Bayes’ Rule and conditional independence assumption to compactly represent the full joint probability distribution using a little space. Given a star pattern $Q$ and Bayesian network $\beta$ learned from table $R$, we have:

\[
sel(Q) = Pr(prop_1 = o_1, prop_2 = o_2, \ldots, prop_n = o_n) \cdot |R| 
\approx Pr_{\beta}(prop_1 = o_1, prop_2 = o_2, \ldots, prop_n = o_n) \cdot |R| 
= \prod_{i=1}^{n} Pr(prop_i = o_i \mid Parents(prop_i) = \bar{o_i}) \cdot |R|
\]

where Parents($prop_i$) denotes the set of immediate predecessors of $prop_i$ in the Bayesian network; $\bar{o_i}$ denotes the set of values of Parents($prop_i$). Note that for computing $Pr(prop_i \mid parents(prop_i) = \bar{o_i})$, we only need to know the values of $prop_i$’s parent properties, which would save a lot of space in practice.

Much research has been conducted in Bayesian network learning. We use K2 [4] algorithm to learn the structure of Bayesian networks.
B. Estimation for chain patterns

The chain graph pattern is another kind of common SPARQL query patterns, which consists a sequence of triple patterns where the object of the previous triple pattern is also the subject of the next pattern. We construct the chain count table TC (shown in Figure 3) for frequent chain patterns, which has two attributes: Head-Chain-Rear and Count. Each row of TC indicates a chain pattern with their frequencies (selectivities). However, chain table TC could be too large to fit in a small amount of main memory.

For a composite graph pattern decomposed into the precomputed patterns, we need to combine the selectivity of precomputed star and chain patterns. There are three basic cases: Case 1 (star-chain join): The composite graph pattern \( Q \) can be decomposed into a precomputed star pattern \( S \) and a chain pattern \( C \) joined on a variable \( Y \). Case 2 (star-star join): \( Q \) can be decomposed into two precomputed star patterns joined on a variable \( Y \). Case 3 (chain-chain join): \( Q \) can be decomposed into into two precomputed chain patterns joined on a variable \( Y \).

For these cases, we go through all values of the join node \( Y \). It is easy to acquire the selectivity of star pattern \( S \) with different values on \( Y \) through inference on the Bayesian network. Similarly, we can obtain the selectivity of chain pattern \( C \) with different values on \( Y \) through the chain histogram. If two patterns have the same value on the join node \( Y \), we combine the selectivity of two patterns in the product form. For the case where a graph pattern \( Q \) can be decomposed into multiple patterns, we first select two joined patterns from \( Q \) and compute the selectivity of the joined patterns. We iterate this process until the overall selectivity of \( Q \) is obtained.

V. Conclusion

In this research, we focus on selectivity estimation for RDF graph patterns. We construct the Bayesian networks and chain histogram for estimating the selectivity of star and chain patterns. For an arbitrary composite graph pattern, we propose the algorithms to combine the results of chain patterns and star patterns that we have precomputed to estimate the overall selectivity.

REFERENCES


Data Management in Scientific Cloud Workflow Systems

Dong Yuan
Faculty of Information and Communication Technologies,
Swinburne University of Technology
Hawthorn, Melbourne, Australia 3122
{dyuan, yyang}@swin.edu.au

Abstract—Data-intensive scientific applications are posing many challenges in distributed computing systems. In the scientific field, the application data are expected to double every year over the next decade and further. With this continuing data explosion, high performance computing systems are needed to store and process data efficiently, and workflow technologies are facilitated to automate these scientific applications. Scientific workflows are typically very complex. They usually have a large number of tasks and need a long time for execution. Running scientific workflow applications usually need not only high performance computing resources but also massive storage.

The emergence of cloud computing technologies offers a new way to develop scientific workflow systems. Scientists can upload their data and launch their applications on the scientific cloud workflow systems from everywhere in the world via the Internet, and they only need to pay for the resources that they use for their applications. As all the data are managed in the cloud, it is easy to share data among scientists. This kind of model is very convenient for users, but remains a big challenge to the system. This seminar will discuss several research topics of data management in scientific cloud workflow systems.

Keywords—data management; scientific workflow; cloud computing;

I. INTRODUCTION

Data-intensive scientific applications are posing many challenges in distributed computing systems. In many scientific research fields, like astronomy, high-energy physics and bio-informatics, scientists need to analyse terabytes of data either from existing data resources or collected from physical devices. During these processes, similar amounts of new data might also be generated as intermediate or final products [9]. According to [20], in the scientific field, the application data are expected to double every year over the next decade and further. With this continuing data explosion, high performance computing systems are needed to store and process data efficiently, and workflow technologies are facilitated to automate these scientific applications. Scientific workflows are typically very complex. They usually have a large number of tasks and need a long time for execution. Running scientific workflow applications usually need not only high performance computing resources but also massive storage [9].

Nowadays, popular scientific workflows are often deployed in grid systems [16] because they have high performance and massive storage. However, building a grid system is extremely expensive and it is normally not open for scientists all over the world. The emergence of cloud computing technologies offers a new way to develop scientific workflow systems.

Since late 2007 the concept of cloud computing was proposed [22], it has been utilised in many areas with some success. Cloud computing is deemed as the next generation of IT platforms that can deliver computing as a kind of utility [8]. Foster et al. made a comprehensive comparison of grid computing and cloud computing [11]. Some features of cloud computing also meet the requirements of scientific workflow systems. Cloud computing systems provide the high performance and massive storage required for scientific applications in the same way as grid systems, but with a lower infrastructure construction cost among many other features, because cloud computing systems are composed of data centres which can be clusters of commodity hardware [22]. Research into doing science and data-intensive applications on the cloud has already commenced [18], such as early experiences like Nimbus [14] and Cumulus [21] projects. The work by Deelman et al. [10] shows that cloud computing offers a cost-effective solution for data-intensive applications, such as scientific workflows [13].

Furthermore, cloud computing systems offer a new model that scientists from all over the world can collaborate and conduct their research together. Cloud computing systems are based on the Internet, and so are the scientific workflow systems deployed in the cloud. Scientists can upload their data and launch their applications on the scientific cloud workflow systems from everywhere in the world via the Internet, and they only need to pay for the resources that they use for their applications. As all the data are managed in the cloud, it is easy to share data among scientists. This kind of model is very convenient for users, but remains a big challenge for data management to scientific cloud workflow systems.

Firstly, new data storage strategy is required in cloud scientific workflow systems. In a cloud computing, theoretically, the system can offer unlimited storage resources. All the application data of the scientific workflows can be stored, including the generated intermediate data, if we are willing to pay for the required resources. Storing all the application data in the cloud is obviously not cost-effective, since some data are seldom used and huge in size. However, in scientific cloud workflow systems, a scientist cannot decide whether a piece of application data should be stored or not, since the data are shared and he is not the only user.

Secondly, new data placement strategy is also required, which means the cloud workflow systems must have the ability to decide where to store the application data. Cloud computing platform contains different cloud service providers with different pricing models, where data transfers between service providers also carry a cost.
The cloud scientific workflows are usually distributed, and the data placement strategy will decide where to store the application data, in order to reduce the total system cost.

Last but not least, new data replication strategy should also be designed for cloud scientific workflow systems. A good replication strategy can not only guarantee the security of application data, but also further reduce the system cost by replicating frequently used data in different locations. Replication strategy in the Cloud should be dynamic based on the application data’s usage rate.

II. RELATED WORKS

As cloud computing has become more and more popular, new data management systems have also appeared, such as Google File System [12] and Hadoop [1]. They all have hidden infrastructures that can store the application data independent of users’ control. Google File System is designed mainly for Web search applications, which are different from workflow applications. Hadoop is a more general distributed file system, which has been used by many companies, such as Amazon and Facebook. When you push a file to a Hadoop File System, it will automatically split this file into chunks and randomly distribute these chunks in a cluster. Furthermore, the Cumulus project [21] introduced a scientific cloud architecture for a data centre. And the Nimbus [14] toolkit can directly turn a cluster into a cloud and it has already been used to build a cloud for scientific applications.

Comparing to the distributed computing systems like cluster and grid, a cloud computing system has a cost benefit [4]. Assunção et al. [5] demonstrate that cloud computing can extend the capacity of clusters with a cost benefit. Using Amazon clouds’ cost model and BOINC volunteer computing middleware, the work in [15] analyses the cost benefit of cloud computing versus grid computing. The idea of doing science on the cloud is not new. Scientific applications have already been introduced to cloud computing systems. In terms of the cost benefit, the work by Deelman et al. [10] also applies Amazon clouds’ cost model and demonstrates that cloud computing offers a cost-effective way to deploy scientific applications. In [13], Hoffa conducted simulations of running an astronomy scientific workflow in cloud and clusters, which shows cloud scientific workflows are cost-effective. The above works mainly focus on the comparison of cloud computing systems and the traditional distributed computing paradigms, which shows that applications running on cloud have cost benefits. When it comes to how to reduce cost of running applications in clouds, Deelman et al. present in [10] that storing some popular intermediate data can save the cost in comparison to always regenerating them from the input data. Furthermore, in [2], Adams proposes a model to represent the trade-off of computation cost and storage cost in storing application data, but the authors have not given the specific method of managing the data.

To the best of our knowledge, research in cost-effectively managing application data of cloud scientific workflow systems is still in blank. My research will focus on this spot and develop novel strategies of data storage, placement and replication in cloud scientific workflow systems.

III. RESEARCH PROBLEMS AND METHODOLOGIES

A. Data storage strategy

1) Problem analysis

Traditionally, scientific workflows are deployed on the high performance computing facilities, such as clusters and grids. Scientific workflows are often complex with huge intermediate data generated during their execution. How to store these intermediate data is normally decided by the scientists who use the scientific workflows. This is because the clusters and grids only serve for certain institutions. The scientists may store the intermediate data that are most valuable to them, based on the storage capacity of the system. However, in many scientific workflow systems, the storage capacities are limited, such as the pulsar searching workflow we introduced. The scientists have to delete all the intermediate data because of the storage limitation. This bottleneck of storage can be avoided if we run scientific workflows in the cloud.

In a cloud computing environment, theoretically, the system can offer unlimited storage resources. All the intermediate data generated by scientific cloud workflows can be stored, if we are willing to pay for the required resources. However, in scientific cloud workflow systems, whether to store intermediate data or not is not an easy decision anymore.

a) All the resources in the cloud carry certain costs, so either storing or generating an intermediate dataset, we have to pay for the resources used. The intermediate datasets vary in size, and have different generation cost and usage rate. Some of them may often be used whilst some others may be not. On one hand, it is most likely not cost effective to store all the intermediate data in the cloud. On the other hand, if we delete them all, regeneration of frequently used intermediate datasets imposes a high computation cost. We need a strategy to balance the generation cost and the storage cost of the intermediate data, in order to reduce the total cost of the scientific cloud workflow system.

b) The scientists can not predict the usage rate of the intermediate data anymore. For a single research group, if the data resources of the applications are only used by its own scientists, the scientists may predict the usage rate of the intermediate data and decide whether to store or delete them. However, the scientific cloud workflow system is not developed for a single scientist or institution, rather, developed for scientists from different institutions to collaborate and share data resources. The users of the system could be anonymous from the Internet. We must have a strategy storing the intermediate data based on the needs of all the users that can reduce the cost of the whole system.

Hence, for scientific cloud workflow systems, we need a strategy that can automatically select and store the most appropriate intermediate datasets. Furthermore, this
strategy should be cost effective that can reduce the total cost of the whole system.

2) **Methodology**

Scientific workflows have many computation and data intensive tasks that will generate many intermediate datasets of considerable size. There are dependencies among different intermediate datasets. Data provenance in workflows is a kind of important metadata, in which the dependencies between datasets are recorded. Data provenance records the information on how the intermediate datasets were generated, which is very important for the scientists. Furthermore, regeneration of the intermediate datasets might start from some stored intermediate datasets instead. In the scientific cloud workflow system, data provenance is recorded while the workflow execution. Taking advantage of data provenance, we can build an IDG based on data provenance. All the intermediate datasets once generated in the system, whether stored or deleted, their references are recorded in the IDG. Based on the IDG, we can calculate the generation cost of every intermediate dataset in the scientific cloud workflows. By comparing the generation cost and storage cost, our strategy can automatically decide whether an intermediate dataset should be stored or deleted in the cloud system, no matter whether this intermediate dataset is a new dataset, regenerated dataset or stored dataset in the system.

Data provenance is important to this strategy in building the IDG. Fortunately, due to the importance of data provenance in scientific applications, much research about recording data provenance of the system has been done. Some of them are especially for scientific workflow systems [6]. Some popular scientific workflow systems, such as Kepler [16], have their own system to record provenance during the workflow execution [3]. In [19], Osterweil et al. present how to generate a Data Derivation Graph (DDG) for the execution of a scientific workflow, where one DDG records the data provenance of one execution. Similar to the DDG, our IDG is also based on the scientific workflow data provenance, but it depicts the dependency relationships of all the intermediate data in the system. Hence, we can build our IDG by taking advantage of these related works.

B. **Data placement strategy**

1) **Problem Analysis**

Scientific applications are data intensive and usually need collaborations of scientists from different institutions [7], hence application data in scientific workflows are usually distributed and very large. When one task needs to process data from different data centres, moving data becomes a challenge. Some application data are too large to be moved efficiently, some may have fixed locations that are not feasible to be moved and some may have to be located at fixed data centres for processing, but these are only one aspect of this challenge. For the application data that are flexible to be moved, we also cannot move them whenever and wherever we want, since in the cloud computing platform, data centres may belong to different cloud service providers that data movement would result in costs. Furthermore, the infrastructure of cloud computing systems is hidden from their users. They just offer the computation and storage resources required by users for their applications. The users do not know the exact physical locations where their data are stored. This kind of model is very convenient for users, but remains a big challenge for data management to scientific cloud workflow systems.

2) **Methodology**

In cloud computing systems, the infrastructure is hidden from users. Hence, for most of the application data, the system will decide where to store them. Dependencies exist among these data. In this paper, we initially adapt the clustering algorithms for data movement based on data dependency. Clustering algorithms have been used in pattern recognition since 1980s, which can classify patterns into groups without supervision. Today they are widely used to process data streams. In many scientific workflow applications, the intermediate data movement is in data stream format and the newly generated data must be moved to the destination in real-time. We adapt the $k$-means clustering algorithm for data placement in scientific cloud workflow systems. Scientific workflows can be very complex, one task might require many datasets for execution; furthermore, one dataset might also be required by many tasks. If some datasets are always used together by many tasks, we say that these datasets are dependant on each other. In our strategy, we try to keep these datasets in one data centre, so that when tasks were scheduled to this data centre, most, if not all, of the data they need are stored locally.

Our data placement strategy has two algorithms, one for the build-time stage and one for the runtime stage of scientific workflows. In the build-time stage algorithm, we construct a dependency matrix for all the application data, which represents the dependencies between all the datasets including the datasets that may have fixed locations. Then we use the BEA algorithm [17] to cluster the matrix and partition it that datasets in every partition are highly dependent upon each other. We distribute the partitions into $k$ data centres, where the partitions have fixed location datasets are also placed in the appropriate data centres. These $k$ data centres are initially as the partitions of the $k$-means algorithm at runtime stage. At runtime, our clustering algorithm deals with the newly generated data that will be needed by other tasks. For every newly generated dataset, we calculate its dependencies with all $k$ data centres, and move the data to the data centre that has the highest dependency with it.

By placing data with their dependencies, our strategy attempts to minimise the total data movement during the execution of workflows. Furthermore, with the pre-allocate of data to other data centres, our strategy can prevent data gathering to one data centre and reduces the
time spent waiting for data by ensuring that relevant data are stored locally.

C. Data replication strategy

1) Problem Analysis

Data replication is also an important issue for cloud workflow systems as presented by the two points below:

First, a good replication strategy can guarantee fast data access for the cloud workflow system. In the scientific workflows with many parallel tasks will simultaneously access the same dataset on one data centre. The limitation of computing capacity and bandwidth in that data centre would be a bottleneck for the whole cloud workflow system. If we have several replicas in different data centres, this bottleneck will be eliminated.

Second, a good replication strategy can reduce data movement between data centres. For example, if tasks in one data centre always need to retrieve data from the same data set in a remote data centre, it is better to replicate that data set in the local data centre to reduce the data movement.

Third, a good replication strategy can guarantee data reliability for the cloud workflow system. Because data centres in cloud workflow systems are built up with massive cheap commodity hardware, the breakdown of some hardware could happen any time. It is essential to keep several copies of each data in different data centres for reliability.

However, at present, data replication strategies that utilised in cloud data management systems are usually static. For example, in Hadoop, users can manually set the number of replicas, and the system will automatically replicate the application data in different places (racks or clusters, depends on the scale of the system). Static replication can guarantee the data reliability, but in cloud environment, different application data have different usage rate, where we should have the dynamic strategy to replicate the application data based on their usage rate.

2) Methodology

The basic strategy for the replication could be as follow:

a) Always keep fix number copies of each dataset in different data centres to guarantee reliability and dynamically add new replicas for each dataset to guarantee data availability.

b) Where to place the replicas is based on data dependency.

c) How many replicas should a dataset have is based on usage rate of this dataset.

REFERENCES


Resource Planning for Business Process

Jiajie Xu
Centre for Complex Software System and Services
Faculty of Information and Communication Technologies
Swinburne University of Technology
Melbourne, Australia
jxu@groupwise.swin.edu.au

Abstract

Resource management is a key issue in business process management as it enables the BPM systems to be optimally executed to each to certain business goals and requirements. A lot of works have been done to investigate this topic in different optimization criterions. In different from existing works, our focus is to plan the optimized resource utilization at the build time, so that the business goals and requirements can be guaranteed before execution. We propose their research problems on this issue, and then present the innovative solutions for them on basis of a planning framework.

Keywords: Business process; Resource management.

1. Introduction

Business process management (BPM) is drawing our attention because the quality of processes is increasingly important for enterprises to thrive or even survive in the fierce market competition nowadays [2]. Lots of efforts has been done in this area [6]. An effective and efficient process would benefit an enterprise with shorter time, better quality and lower cost. Resource management is a key in the area of BPM as it enables the resources of enterprises to be well used. Given that resource allocation must comply with certain requirements like from different, it is essential to plan the use of resource at the build time, so that enterprise resources can be rationally used to guarantee the business requirements can be satisfied before execution. Also, some necessary information can be provided for decision enterprise decision making for customer negotiations if their requirements cannot be satisfied by available resources.

We reckon that resource allocation and business process impact on each other. Structures of business process set a constraint on how resources are allocated to tasks due to the dependency. However, it is possible that a business process is not well-defined, and as a result the resources may not be utilised optimally to reach certain business goal. It is desired that the structure of a business process is improved so that resources can be utilised in a more optimal way. However, as far as we know, no work has discussed this kind of improvement. In this paper, we collectively discuss the problems of resource allocation optimisation for business processes, and resource oriented business process improvement.

The resource planning can be further complicated in the scenario where a massive number of process instances exist. The dependencies and conflicts between concurrent process instances and exclusively occupied resources have to be taken into account in the resource planning. For an organisation, resources must be allocated to guarantee the assigned volume of process instances can be finished before a negotiable deadline to customers at an acceptable lower cost to organisation. Due to those factors such as inventory capacity and resource availability patterns, the massive number of process instances have to be executed in batches for balancing the load.

Also, resource availability set a constraint on how resources can be used [3, 4]. Resource allocation must comply with such constraint, and meet the requirements of customers as much as possible. For example, doctors in the hospitals are available on the duty list. The hospital wishes to use the resources correctly and optimally for certain business goals. We try to handle this problem by considering availability patterns of resources in the process instance scheduling.

The remainder of this paper is organized as follows. Section 2 represents a basic framework for the research issue to be investigated. Section 3 introduce three problems and outline the solution on each of them. Finally, Section 4 addresses the conclusion of this paper.

2. Framework

The framework in Fig.1 illustrates the relationship of some notions including resources, roles, tasks and business process. A business process consists of a set of tasks and gateways. Each resource performs as one role, and one role may represent multiple resources. Any task can be executed by a set of roles, and a role may be capable of executing many tasks. When allocating resources to tasks, the allocation is subject to the dependency due to the structure of the process and the roles of resources. Cost is an attribute of a role, and it denotes the money that the enterprise has to pay for resources of that role when they are allocated to execute tasks. Time for a task to be executed is determined by
which role is assigned to perform this task. In our framework, one resource can serve for one task at most at one time. Otherwise, we say it is an allocation conflict.

Fig. 1. Framework

Based on this Framework, resource planning for business processes in this paper is investigated on the following three problems in different scenarios for application.

3. Key Research Problems

Based on the issue motioned above, we introduce three problems related to this issue, and introduce the methodologies of solving these issues. The first problem is to integrate resource planning together with process change management, to enable them adaptive with each other. The second

3.1. Resource allocation with process change management

As we know, the process structure set a constraint on resource planning. On the other side, the business process may not be well defined, and it will cause the resources unable to be optimally utilized to reach certain business goals. As such, it is important to improve process structure along with resource allocation to support optimal resource utilization. That is to say, it is necessary to combine resource allocation together with process change management in process execution planning due to their impact on each other.

As the cost and time for executing a task are unknown until it is allocated with actual resource, the analysis on the business process performance is inevitably involved with resource allocation. Resource allocation to tasks will be done in such an optimised way that cost is minimal while satisfying time constraint. Based on a set of heuristic rules [1], the optimised resource allocation for business process is carried out by the following three allocation steps:

(1) A basic allocation strategy will be applied to searching for a resource allocation satisfying Rule 3 and Rule 1, which aims the minimal expense for executing a business process with balanced allocation for all paths. Surely, no resource is allocated to more than one task at any time.

(2) In case that the allocation scheme in Step (1) violates Rule 2, an adjustment strategy will be applied to shorten the execution time by re-allocating resources until time constraint is satisfied.

(3) In case that the time is less than the limit from Step (1) or Step (2), according to Rule3, the adjustment strategy will be applied to do the resource oriented business process improvement in order to achieve a lower expense while maintaining the time constraint to be satisfied.

Four algorithms are listed in work [1] for the three step resource allocation.

3.2. Resource planning for massive number of process instances

One challenge for resource planning is, when the number of process instance is very large, the complexity is huge as due to the inter dependencies of process structure and the intra dependencies of different instances. Also, resources planning for process instances must consider a deadline posed by customers, and a lower cost, conflicts in concurrent use of resources, periodical resource availability according to human nature and inventory capacity, etc. from the organisation point of view.

A fundamental approach in [5] called holistic solution is applied to plan resources for all instances. In the holistic planning strategy of this paper, we propose a two step solution based on a set of heuristic rules. The heuristic rules are given as the instance Resource allocation should avoid, if possible, increasing the overall time of the longest path, while it tends to intends to use economical resource as task performer in order for minimal overall cost if the longest path is not increased. In the first step, a basic resource allocation is applied to plan the resources for all instances for build time optimisation. However, due to the process structural characteristic and instance dependency, run time conflicts may occur in the first step. Therefore immediately after, we handle those conflicts by adjusting previous allocation as second step.

Resource planning can be also carried out by the batched way, where instances are executed in batches. The batched processing of process instances can improve the planning efficiency and support resources periodically used to cater for their nature. We will address the problem of how to partition instances into batches, and then optimise the use of resource according to the fixed batch pattern. First of all, we conduct an analysis on available resources, business requirements and process structures to sketch out a preliminary batch pattern for the resource planning. The default batch pattern is selected by applying:

(a) resource checking; (b) lower bound update; (c) batch pattern update recursively. Based on the default batch pattern, resource allocation is made as the holistic approach. If the given deadline cannot be satisfied, we update the batch pattern. This procedure continues until a
valid allocation scheme is found or cannot find an allocation plan for even a single batch.

3.3. Process scheduling based on resource availability patterns

In some specific applications, resources may have certain patterns, and resource planning should follow those resource patterns. It is important to generalise some basic resource patterns, and investigate how to manage and plan the resources under those patterns. Also, business process has some structural patterns, and the structural patterns have impact on resource planning. What are the patterns for resources and processes, and how resource and process structure management should be adapted to the patterns for optimisation?

Other than resource allocation, we schedule the process instances to resources instead of the traditional way. Resources have availability patterns on it, and process instances have a number of tasks in certain dependency determined by the process structure. Such instance scheduling is applied on basis of some heuristic rules. We proposed three strategies in different optimization criterion. Further, the strategies are compared and evaluated in different application scenarios by an experimental study.

4. Discussion

This paper discussed the problem of resource allocation for business processes. An approach was proposed to allocate resources to tasks in such a way that the total expense is minimal while the requirement of executing time on a business process is satisfied. In this approach, a basic strategy is applied first to minimise total expense. Then, an adjustment strategy is applied to modify allocation such that the time constraint is met in a smart way. The advantage of this approach over previous approaches lies in the relationship between the effective resource allocation and the business process improvement. To cater for the resource allocation requirements and available resources of an enterprise, the structure of a business process can be changed. After the structure of the business process is changed, the performance of the business process in terms of better utilising resources is improved.

Also, we investigated how to plan resources for a business process optimally, in aspect of meeting process requirements and rational utilisation of resources. Also, our approach can provide information to organisation for decision making and negotiation with customer in order to better use resources. When a massive number of instances are given, both inter and intra instance dependencies are considered in the planning of each batch of instances to be executed in parallel. As the problem is computationally hard, a set of heuristic rules were designed, and two strategies based on these heuristic rules were proposed and compared.

References

Parameter control in evolutionary algorithms

Aldeida Aleti
Faculty of ICT, Swinburne University of Technology, Australia
aaleti@swin.edu.au

Abstract—Deciding on the best performing parameter setting for evolutionary algorithms in a problem domain is a non-trivial task. For example, in the case of the deployment optimisation of software components in automotive systems, despite the wide range of evolutionary algorithms already published, it is still unknown which parameter settings of the algorithms are the optimal choices. The values of the parameters of an evolutionary algorithm are dependent from the given problem instance and greatly determine whether the algorithm will efficiently converge to the optimal solutions or not.

The selection of the parameter values is, however, a difficult task. The existing approaches for setting the parameter values of evolutionary algorithms follow two main streams, those who set the parameter values in advance of the search process, i.e. parameter tuning, and those who change them during the search process, referred to as parameter control. Optimal parameter tuning is not only hard and time consuming but also impossible, since different values of the parameters are optimal at different stages of the search. On the other hand, although parameter control is more efficient than parameter tuning, choosing a well-working procedure among the different possibilities one can design is also a hard task. The approach proposed will use statistical information to design an efficient parameter control procedure which overcomes the drawbacks of previous works done in this field. To validate the approach, multiple runs of the optimisation algorithms on realistic case studies taken from the automotive industry will be considered.

Keywords-component deployment; optimisation; parameter setting;

I. INTRODUCTION

Design decisions that map software components to hardware hosts have an important impact on the quality of the resulting embedded system [15], [11]. Due to the complexity of these systems, the best design options have to be selected from a combinatorial growing design space. Since this problem is known to be NP hard, approximate methods should be employed, which find acceptably good solutions in a fixed amount of time [5], [18]. Moreover, the component deployment problem often involves concurrent optimization of several incommensurable and competing quality objectives. The solution to such problems is usually a set of design alternatives, which assures a tradeoff between the conflicting qualities, referred to as Pareto-optimal solution set or briefly Pareto front. The employment of approximate methods for multiobjective optimisation yields approximations of the Pareto front, i.e. approximate sets.

Many approximate optimisation methods exist in the literature for the component optimisation problem [16], [17], [19], [20], [15]. Evolutionary algorithms have proven to be powerful and robust in optimising problems where little or no domain knowledge is inaccessible [2]. Although researchers generally agree on using evolutionary algorithms for component deployment optimisation, no attention is given to the parameter configurations of the algorithms. As Blum [4] states, the performance of an approximate optimisation method is highly dependent on the context in which it is being used, requiring specific parameter settings of these algorithms.

The issue of setting the values of various parameters of approximate methods is crucial for achieving a good performance. The existing approaches for parameter setting can be divided in two groups: those who set the parameter values in advance of the search process [8], [10], also called parameter tuning and those who change them during the search process, referred to as parameter control [7], [21].

Setting the parameters in advance may cause over confidence or over tuning of the parameters. This happens when an algorithm is tuned based on a finite number of training instances causing cost estimates to be orders of magnitude too optimistic, which leads to strongly impaired performance [9]. On the other hand, when there are few parameters, parameter tuning tends to be easy. In this case it is often convenient to try each combination of the parameter values. This process is called full factorial design which is an automatic algorithm configuration based on local search [9]. If an algorithm has too many parameters, one often reverts to local optimisation. Full factorial design becomes intractable in such situations, since the number of possible configurations grows exponentially with the number of parameters.

On the other hand, designing an optimal parameter control method is hard. The existing parameter control methods can be divided in three main groups: deterministic, adaptive and self-adaptive parameter control [9]. In deterministic parameter control, static parameters are tuned over time. Tuning static parameters is not only hard, but sometimes impossible [9]. Adaptive methods use information about the current state of the search. This raw information needs to cleverly be interpreted and used to update the parameters. These methods can only be as good as the information that they get and the update rules that they perform according to this information. The last group of parameter control is
called self-adaptive methods and relies on the selection step of the optimisation method. Either an optimisation algorithm is used to find the optimal parameter configurations or the parameters to be configured are encoded in the solutions, so that not only the fittest solutions but also the individuals with the best parameters survive.

To conclude, in the existing methods for parameter setting of evolutionary algorithms the user needs to define which parameters need to be tuned. Moreover the rules used to change the values of the parameters need to be defined before the search process. In the case of stochastic methods, such as evolutionary algorithms, the effect of the parameters and their correlation is not known beforehand. Therefore, the selection of parameters to be tuned and the rules for changing them should be determined during the search. Accordingly, the method proposed in this work will use evidence from the search process to decide upon the parameters that have an effect on the search, their correlations and the rules that should be applied to change them.

II. RELATED WORK
A. Optimization Methods in Component Deployment

The different methods used to help the designers find an optimal deployment of software components on the hardware platform can be classified into two main groups: exact methods and approximate methods. Exact algorithms are guaranteed to find an optimal solution and to prove its optimality for every instance of a problem. The run-time, however, increases with the instance size, and often only small sized instances can be practically solved to provable optimality. Consequently, when we have larger instances we need to trade optimality for run-time, yielding approximate algorithms. In other words, the guarantee of finding optimal solutions is sacrificed for the sake of getting good solutions in a limited time.

Exact methods such as Binary Integer Programming (BIP) [3] or Graph Cutting [12] have been applied to component deployment optimisation. However, in [12], the authors prove that finding solutions for an unconstrained deployment problem is NP-hard. Even when applying constraints that help on restricting the possible search space, exact algorithms were found to be suitable only for small instances of the problem [14].

Approximate methods used in component deployment problem can be divided into two main categories, i.e. iterative algorithms, used in [16], [17], [13], [19], [20], [15], and constructive algorithms proposed by Csorba [6].

Iterative algorithms generate new candidate solutions using only the current solution or the current population of solutions. Typical representatives of this class are Evolutionary Algorithms [16], [17], [19], [20], [15]. The main problems with these algorithms are the time complexity for the convergence to the real Pareto-optimal set and their proneness to confinement in local optima. The main reason for this premature convergence is the wrong setting of the parameter values. While iterative algorithms have been thoroughly investigated and are relatively well understood, constructive algorithms, another subclass of approximate methods are still not widely applied. The work of Csorba et al. [6] is the only approach that applies a constructive algorithm in our problem domain. The approach uses Cross Entropy Ant System (CEAS), which belongs to the class of Ant Colony Optimisation algorithms.

B. Parameter Setting in Evolutionary Algorithms

We can distinguish two different ways of setting parameter values: parameter tuning [8], [10] and parameter control [7], [21]. Parameter tuning means that good values for the parameters are found before running the algorithm and then used for running the algorithm. In this case the parameters are static, which means that they don't change during the search process. Alternatively, parameter control starts a run with non-optimal parameter values which evolve to the optimal parameter values during the search process. However, there is little information about which parameter setting method is more effective in certain circumstances.

III. RESEARCH QUESTION AND APPROACH

The research questions addressed in this work are as follows:

- Which are the appropriate model and rules for the parameter control of evolutionary algorithms?
- How can the current data observed in the search process be used for parameter control?
- Since only local data is being used, how can we understand that this local tuning is accurate?
- How can the effect of each parameter and the interaction between the parameters be assessed?
- How can we predict the effect the change in the parameter values will have in the performance of the algorithm?

The premise of the approach taken is that a precise tuning of the parameter values is not only time consuming but also impossible, due to the changing behaviour of the approximate optimisation methods during the search process. Similarly, the current parameter control techniques need a lot of effort and are not optimal solutions for the parameter setting, since the parameters to be tuned are selected prior to the start of the search process. Knowing the effect of parameters and their interactions beforehand is impossible, therefore deciding on which parameters to tune and the tuning rules is a hard job. Rather, some knowledge of the current search process may suffice to predict which are the parameters that need to be changed, how should the new values be tuned and which will be their effect on the future search process.

In order to address the above problems, evidence of the observations will be used, which will not only effect the
update rules of the parameter values, but also will help in
deciding which parameters need to be changed. Due to the
stochastic nature of the evolutionary algorithms, Bayesian
inference will be used. Bayesian inference uses collected
evidence, which may be consistent or inconsistent with a
given hypothesis, to infer the probability that a hypothesis
may be true. The accumulated evidence changes the degree
of belief in a hypothesis during the search.

The validation of the approach will be achieved by
multiple runs of the optimisation algorithms on realistic
case studies taken from the automotive industry. Since the
approximate algorithms generate data drawn from a normal
probability distribution, a parametric testing technique will
be used. As an experimental framework the Archeopterix [1]
tool will be employed.

IV. CONCLUSION

In this work, different methods for parameter setting
in evolutionary algorithms will be investigated and a new
method for parameter control will be developed. The new
approach to parameter control will be focused on statistical
data from the search process of the evolutionary algorithms,
which will guide to better selection of the parameters to be
changed and more efficient update rules based on the effect
of the parameters and their interactions.

ACKNOWLEDGMENT

The author would like to thank dr. Lars Grunske for the
support that made this work possible.

REFERENCES

ArcheOpterix: An extendable tool for architecture optimiza-
tion of AADL models. In MOMPES’09, pages 61–71. IEEE

tionary optimization. In Evolutionary Programming, pages
433–439, 1996.

A binary integer programming model for optimal object

optimization: Overview and conceptual comparison. ACM

International Conference on Software Engineering (ICSE’06),

deployment of collaborating components. In DAFS, volume


of Genetic Adaptive Systems. PhD thesis, University of
Michigan, 1975.

Parameter control in evolutionary algorithms. In Parameter
Setting in Evolutionary Algorithms, volume 54 of Studies in

genetic algorithms. IEEE Transactions on Systems, Man, and

systems - a generic framework. Journal of Systems and

partitioning system. In OSDI: Operating System Design and

for the design of integrated dependable real-time embedded
systems. In EUC, volume 4808 of LNCS, pages 517–530.

[14] T. Kichkaylo, A. Ivan, and V. Karamcheti. Constrained com-
ponent deployment in wide-area networks using AI planning
techniques. In IPDPS’03), page 3. IEEE Computer Society,
2003.

[15] S. Malek, M. Mikic-Rakic, and N. Medvidovic. A decentral-
ized redeployment algorithm for improving the availability of
distributed systems. In Comp. Dep., volume 3798 of LNCS,

model-based automated safety analysis and optimisation tech-
niques. The Journal of Systems and Software, 76(1):77–89,
2005.

approach to exploring embedded system architectures at
multiple abstraction levels. IEEE Transaction Computers,

engineering for automotive systems: Aroadmap. In Interna-
tional Conference on Software Engineering,ISCE’07, pages

[19] D. Salazar, C. M. Rocco, and B. J. Galvan. Optimization of
constrained multiple-objective reliability problems using
evolutionary algorithms. Reliability Engineering & System

with uncertainty of combining randomness and fuzziness.
European Journal of Operational Research, 157(3):716–735,
2004.

multi-objective evolutionary algorithm based on grids. IPL:
Towards a more accurate Census
Integrated Intelligent Decision Support for Field Design and Management of Census Operations in Australia

John D Moore
Director, Census Evaluation Section
Australian Bureau of Statistics
Canberra, Australia
john.moore@abs.gov.au

Abstract—The Census of Population and Housing is a large and complex operation carried out every five years in Australia. In order to improve Census data quality, a virtual environment is being developed to test Census participation strategies. The simulations developed predict non-response for the 2011 Census and help to minimise non-response for the 2016 Census.

Keywords—Census, data mining, clustering, data modelling, agent-based simulation

I. INTRODUCTION

Every 5 years the Australian Bureau of Statistics (ABS) conducts a Census of Population and Housing. The objective of the Census is to accurately measure the number and key characteristics of people in Australia on Census Night [1]. Conducting a Census is one of Australia’s largest piece-time operations, employing over 40,000 temporary staff with a budget of $300 million in 2006, and a budget of $400 million for 2011.

To run a Census the ABS divides Australia into small geographic areas called Collection Districts and assigns Collectors to distribute and collect the Census forms which are filled out by the Australian population. In 2006 collection could be via the internet, an option taken up by around 10% of the population. In 2006 there were 38,704 collection districts and around 35,000 collectors employed.

Some groups of people in the population are undercounted in the Census. These include Aboriginal and Torres Strait Islander people, ethnic groups who have trouble reading or speaking English, the homeless and people with certain disabling conditions which prevent them from filling in a Census form. In addition, some areas are more difficult to enumerate, including secure apartment buildings and remote communities. Special strategies have been developed by the ABS to ensure a more complete count of these groups and areas.

II. THE PROBLEM

The problem is that the number of occupied dwellings that do not send in a completed Census form has been steadily rising over the last three Censuses; as measured by the rising rate of dwelling non-response: 1.8% in 1996, 2.4% in 2001 and 4.2% in 2006 [2]. In response to this increase the Australian Research Council (ARC) and the ABS set aside funds for the “Integrated Intelligent Decision Support for Field Design and Management of Census Operations in Australia” project under the management of DVC(Research) Andrew Flitman of Swinburne University.

III. THE PROJECT

The aim of the project is to provide a facility to enable testing of non-response mitigation strategies to ensure a complete count using models, rather than expensive real-world field tests. The hypotheses being tested in this project include:

1) That modern data mining techniques can be used to discover non-trivial dependencies between various factors influencing census response rate and bias.

2) That an integrated simulation modelling approach that combines the power of both system dynamics and agent-based simulation can be used to create a decision support model for testing the effect of management interventions within survey operations context

3) That an integrated intelligent decision support tool can positively affect field design and management of survey operations.

The project has the following stages:

- developing a process-based domain-specific ontology that describes conceptual relationships between factors influencing census response rate and bias;
- quantifying these relationships using data mining;
- supplementing information on these relationships using qualitative studies of people in areas of low Census response; and
- dynamic agent-based simulation of interactions between these factors for the purposes of prospective evaluation of policies and sensitivity analyses.

IV. PROJECT APPROACH

This approach is similar to one undertaken by the Office of National Statistics in the United Kingdom in Conjunction with the University of Southampton. Models were developed for testing enumeration strategies [4] and to predict best times for conducting interviews [5]. The applicability of
these models to Australian circumstances is limited due to differences in the Census field operations of Australian and UK, and the differences in geography and demographics between the two countries. The ABS has also conducted in-house analysis and noted the effects of collector age and dwelling structure on non-response rates [6].

The approach taken to develop a process-based domain specific ontology was to start with the Business Process Maps developed by the Technology Applications area within the ABS that built the ABS Census systems, together with the Data Dictionary that was used to describe the variables. Converting these two sources into an OWL-based ontology is about one-third complete.

V. DATA MINING

The data mining process has been the bulk of the effort of the PhD project to date. In May 2008 data was gathered from many sources within the ABS: the data included Census output data (official Census data published on the ABS website), metadata (data that describes the output data) and paradata (data that describes the process of gathering, transforming and publishing the output data). Connecting this data and storing it in a manner that enables data mining is a time-consuming and non-trivial task, given the data volumes and confidentiality requirements. The data organisation process is ongoing as different analytical techniques demand data to be organised in different ways.

The data analysis phase of the project was made up of four broad activities: early frequency analysis revealed simple relationships between variables and non-response; chi squared analysis ranked these relationships; Classification and Regression Tree (CART) analysis re-ranked these variables taking into account ‘covariance’ (the influence of one variable on another); finally, clustering was used to segment the data into groups (or clusters) that are meaningful to Census managers and assist in refining non-response strategies.

A. Data Mining Methods

The first method, frequency analysis, involved calculating the non-response frequency for each value of a categorical independent variable. If the frequency of non-response changed for any given value, then that variable could be considered as having some association with non-response.

The next step in the analysis was to develop this idea further using a chi-squared test of independence [8] to determine if any variable had a significant ($\alpha = 0.01$) association with the non-response target variable. The “Cramer’s V” statistic was used to rank variables by the strength of association with non-responses.

Classification and Regression Tree (CART) analysis is a non-parametric (not reliant on the assumption that the data is drawn from a given probability distribution) technique that, in this case, produced a classification tree with nodes split on how well a variables value can differentiate observations based on the dependant variable; non-response [3]. The method used in this analysis was to ‘grow’ the tree from a sample of 40% of the observations and use a separate sample of 30% of the observations to “prune” the tree back to those nodes with a misclassification rate of less than 5%.

B. Data Mining Results

Detailed results from the Data Mining process are under embargo by the ABS until corrective strategies are in place that will mitigate against any possible loss of confidence in Census data in the lead up to the 2011 Census.

What can be discussed is the distribution of non-response from the 2006 Census. The Ward clustering algorithm was used, which is an agglomerative clustering method which computes the distance between two clusters as the ANOVA sum of squares between the two clusters summed over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from previous generation [7]. The clustering process created the segments displayed in Table 1 which describes the non-response rates for the four clusters or “segments”. Segment number four is the thirty eight collection districts (CDs) with high levels of non-responding dwellings. In Australia the average number of dwellings per collection district was 225 (ABS, 2006). The average number of non-responding dwellings in these CDs was 207.14. The 355 CDs in Segment number three have also got a high number of non-responding dwellings, on average 81.4 dwellings per CD. Segment number two contains 3825 CDs with an average number of non-responding dwellings of 27.67. Segment one holds by far the bulk of Collection Districts – 33,825 of them – with a low average number of non-responding dwellings per CD. The last two columns of Table 1 show the proportions of the segments to the overall number of dwellings in Australia, and to the proportion of the non-response. Segments four, three and two have a disproportional high non-response level.

<table>
<thead>
<tr>
<th>Segment Number</th>
<th>Selection Rule</th>
<th>Number of CDs selected</th>
<th>Mean number of non-responses</th>
<th>Proportion of total dwelling population</th>
<th>Proportion of total dwelling non-response</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>NRC &gt;= 141.5</td>
<td>38</td>
<td>207.14</td>
<td>0.20%</td>
<td>2.40%</td>
</tr>
<tr>
<td>3</td>
<td>56.5 &lt; NRC &lt; 141.5</td>
<td>355</td>
<td>81.14</td>
<td>1.45%</td>
<td>8.99%</td>
</tr>
<tr>
<td>2</td>
<td>17.5 &lt; NRC &lt; 56.5</td>
<td>3825</td>
<td>27.67</td>
<td>12.77%</td>
<td>33.32%</td>
</tr>
<tr>
<td>1</td>
<td>NRC &lt; 17.5</td>
<td>33,825</td>
<td>5.19</td>
<td>85.57%</td>
<td>55.29%</td>
</tr>
</tbody>
</table>

Note: NRC = Non Response Count = the number of non-responding occupied private dwellings within the CD

69
Segmenting the data in this manner has enabled Census management to focus their energy on the most significant contributors to non-response.

The data analysis phase is approaching its conclusion. Tasks left to complete are aggregating the data to ‘Collector Workloads’ rather than collection districts and to add collector salary information to the database of information available for data mining. This database will be made available for ABS staff for further analysis work. The main output of the analysis phase will be a list of ranked variables predicting non-response with corresponding parameter estimates. In the mean time, reports describing factors and their relationships, and comparisons between various geographic areas within Australia have been produced for an ABS audience in order to improve planning for the 2011 Census.

VI. QUALITATIVE RESEARCH

Qualitative analysis is being undertaken by Dr Meg Carter of the Institute of Social Research at Swinburne University. Dr Carter has published an overview of her methodology and expected outputs on sisr.net

VII. AGENT BASED SIMULATION

Agent based simulation work has commenced on three fronts. The first is to develop an Australian version of the Census enumerator model developed by Simon Doherty of the University of Southampton. The main changes that are necessary to the Doherty model are changing the parameters to the underlying algorithms (based on data collected in the data analysis phase) and changing the interventions available to Australian Census management. This work is close to completion.

The second strand of simulation work is the development of statistical models that explain non-response. The first of these is a model that predicts non-response in a particular Collection District based on the expected rate of contact between collector and respondent. This simple model will serve as an enabler for more complex models, in particular a model that predicts non-response rate based on instructions to collectors on the number of times they should attempt to visit a house on the form delivery and form pick-up phases. This work is in its infancy.

The last strand of simulation work is to develop a graphical interface to the models to form an Integrated Intelligent Decision Support system that predicts non-response for the 2011 Census and allows for optimising response rates for the 2016 Census. This work is in its early days; evaluation of graphical modelling systems is taking place.

The PhD project has used 58% of its allocated time, and is estimated to be 50% complete. This is not a problem as the next phase of the project is IT related which is the background of the PhD candidate.

The simulation system will be validated by comparison with the results of the “2011 Census Major Test”, scheduled for July 2010.

ACKNOWLEDGMENT

Thanks to the many Census staff at the ABS in the Field, Payroll, Output and Processing teams who freely gave their time to provide an understanding of the Australian Census. Thanks also to Martina Hoffmann (ANU) for her valuable statistical advice.

REFERENCES

Multistage Fuzzy Decision Making in Bilateral Agent Negotiation

Jan Richter
Centre for Complex Software Systems and Services
Swinburne University of Technology
Melbourne, Australia
Email: jrichter@groupwise.swin.edu.au

Abstract—This paper presents a multistage fuzzy decision approach to model adaptive strategies in agent-based negotiation. The process of negotiation is thereby considered as an optimization problem of finding the best course of actions leading to a desired outcome given the limited and uncertain information about the negotiation partner’s behaviour and the environment. It is shown that by using fuzzy state transitions for representing the uncertain knowledge about the opponent’s concession behaviour and fuzzy constraints for the agent’s strategy, adaptive strategies can be obtained by only using two reference cases. The proposed approach is different to traditional methods in the sense that it optimizes towards a predefined fuzzy goal providing a more flexible preference model in terms of utility maximization.

Keywords—agent; negotiation; strategy; multistage; fuzzy; decision; qualitative;

I. INTRODUCTION

Negotiation is a process of decentralized decision making between two or more parties who seek to find an agreement satisfying their requirements in the presence of limited common knowledge and conflicting goals [1]. In that sense, it is unique due to the conflict all involved parties find themselves in as they cooperate and compete at the same time [2]. In multi-agent systems, it is therefore considered key to facilitate intelligent decision making and conflict resolution between autonomous software agents. Automated negotiation has also been studied in the research area of artificial intelligence due to its potential in many real world applications including electronic commerce [3], resource allocation and scheduling, task distribution [1] and service-oriented computing [4]. With the aim to negotiate effectively and achieve optimal outcomes a large number of learning and reasoning mechanisms have been studied ranging from simple If-then rules to more advanced learning and reasoning techniques [1], such as Bayesian reasoning [5], genetic algorithms [6], reinforcement learning [7], neural networks [8] or non-linear regression analysis [9]. However, in most cases the agents are required to have either prior knowledge in the form of empirical data or some of the opponent’s parameters, to explore their environment to gather the necessary data or to make assumptions about the underlying decision models the opponent may choose from. For example, the widely used method of case-based reasoning uses past successful interactions to negotiate similar agreements but in most scenarios adjusts only the parameters of the decision model [6] thereby inhibiting the use of cases by agents with different decision models. Relatively few efforts have been reported about using Markov decision processes for modelling negotiation strategies. Narayanan and Jennings [10] model the agent’s behaviour by defining the states in terms of resource availability, deadlines and reservation values where counteroffers are proposed based on the opponent’s offers and changes in those three realms. Even though agreements are achieved much faster when both agents use this algorithm no results are provided for cases where only one agent uses such a strategy. Teuteberg [11] models the behaviour of the opponent by using a probabilistic approach to generate the transition matrix based on a predefined set of opponent tactics. However, a large number of negotiations is needed to obtain sufficient empirical data for reliable state transitions. Negotiation has also been modelled as a fuzzy constraint satisfaction problem in [12] where constraints, preferences and objectives are represented uniformly as fuzzy sets which are distributed among the agents and iteratively relaxed during the exchange of offers [1]. The search process is guided by ordering and pruning the search space but still requires particular negotiation strategies for proposing offers [13]. In the above approaches the major aim is to develop negotiation strategies and mechanisms which are able to adapt to the behaviour of the opponent in order to achieve better outcomes in a distributed environment while using only limited and uncertain information about the other agents. Many of the assumptions, however, appear to be difficult to fulfil due to the dynamic nature of the entire system, the unknown parameters of negotiation partners or the long learning times the agent’s need to explore and obtain the knowledge. The approach presented in this paper models the negotiation process using multistage fuzzy decision making where negotiation is regarded as an optimization problem of finding the best course of actions leading to a desired outcome. The agent’s preferred strategy is represented by fuzzy constraints over its action space whereas the uncertain and limited knowledge about the negotiation partner is modelled using a fuzzy Markov transition matrix enabling qualitative decisions at each stage during the encounter via fuzzy dynamic programming. The preference over the final outcome is specified by a fuzzy goal which directly influences the proposed course of actions. The advantage of this approach is that the available knowledge or beliefs as well as the preferred strategy and outcomes are
specified using fuzzy sets allowing qualitative reasoning for proposing counteroffers. This enables the application in many real world scenarios where information is imprecise and soft.

The reminder of the paper is structured as follows. The next section recalls the basic concept of multistage fuzzy decision making, whereas Section III presents the modelling approach for negotiation strategies and Section IV some initial experimental results. The stage in the PhD research project is outlined in Section V, and, Section VI concludes the paper.

II. MULTISTAGE FUZZY DECISION MAKING

In what follows we present the basic idea of multistage fuzzy decision making in fuzzy environments and refer to [14], [15] for more details. Suppose we have a system with a finite set of crisp states those underlying dynamics are represented by a transition function \( f \), changing the state \( x_t \) of the system at time \( t \) given an action \( u_t \). With the aim to find the best course of actions transforming the system into a desired state, we further assume that the decision process is conditioned on a soft and imprecise environment represented by a fuzzy goal \( G^N \) at the last stage \( N \) and fuzzy constraints \( C^t \) imposed at each individual stage (such a decision process is shown in figure 1). The type of system may be thereby different depending on whether it is deterministic, stochastic or fuzzy in nature. Given the transition function and the fuzzy constraints the aim is now to optimize the course of actions which achieve the best outcome in terms of the predefined fuzzy goal. Solution candidates are obtained via the method of fuzzy dynamic programming which uses recurrence equations [14] to generate an action policy (proposing an action for each state) in each stage starting with the fuzzy goal and iterating backwards while applying the individual fuzzy constraints. Even though this technique is the most widely used method for such problems due to its efficiency and simple implementation, a wide range of different algorithms have been proposed for different variants of the model [15], such as branch-and-bound, neural network or genetic algorithms for deterministic systems; iterative or graph-theoretic approaches for systems with implicitly specified termination times; or interpolative reasoning for systems with fuzzy states and actions.

III. MODELLING APPROACH

To model the negotiation process as a multistage fuzzy decision problem we need to specify the the underlying negotiation mechanism, the state and action space as well as the dynamics and fuzzy environment of the model:

1) Bilateral Negotiation: The prominent model for bilateral negotiation is the service-oriented negotiation model introduced by Faratin et al [16] where two agents propose offers and counteroffers alternately on a number of issues such as price or delivery time. For each issue under negotiation, each agent has a negotiation interval defined by its initial and reservation value; if the intervals of both agents overlap an agreement is generally possible. The agent’s have a preference model which is typically expressed through utility functions scoring the exchanged offers in their respective negotiation intervals and may include other factors such as discounts or negotiation costs. The negotiation ends when one agent accepts or withdraws from the encounter. A withdrawal may occur if the agent reaches its deadline or the agreement is no longer needed due to outside factors. Typically, an offer is accepted if the utility of the current opponent’s offer is equal or higher than the utility of the agent’s next counteroffer.

2) States and Actions: The state and action space of the agent correspond to the opponent’s and its own offers during the encounter. The agent creates the state and action space with the first offer proposals by discretizing the negotiation range and its agent’s negotiation interval.

3) Fuzzy state transitions: The transition matrix of an individual agent encodes its uncertain knowledge about the opponent’s strategic concession behaviour and its responses that may lead to an agreement. Because of the uncertainty inherent to agent negotiation the state transitions are expressed by fuzzy relations. The fuzzy transition matrix may then be derived from limited knowledge about past experiences, the domain in which negotiation takes place or the beliefs of the agent. In our ongoing work we focus on the scenario where only two reference cases (e.g. from past interactions) are available, and their similarity is used to create and update the agent’s fuzzy transition matrix during the negotiation encounter.

4) Fuzzy Goal: The fuzzy goal represents the agent’s preference over the state space, i.e. the opponent’s offers. Due to the qualitative nature of the model, the goal specifies a preference ordering rather than a particular utility value. In the context of negotiation, the degree of membership in the fuzzy goal increases for states closer to the initial value of the agent as they are more preferable to states close to the initial offer of the opponent.

5) Fuzzy Constraints: The fuzzy constraints are the agent’s soft preferences over its action space and thus represent its own concession behaviour during the encounter. The constraints can be time- or behaviour dependent, thereby the effect of the constraints depends to a large degree on the support and the area of the fuzzy constraints. In general, the larger the support and area the stronger the influence of the cases on the actions and vice versa. The fuzzy constraints basically allow the agent to mix its soft preferences with the uncertain knowledge about the opponent’s concession behaviour.

---

Fig. 1. Multistage Fuzzy Decision Process

![Multistage Fuzzy Decision Process](image-url)
IV. Evaluation Results for the Model with Finite Termination Times

The proposed modelling approach has been evaluated in [17] in a single-issue negotiation scenario where both parties may have only partial overlap of negotiation intervals and different negotiation deadlines (for the experiment settings see [17]). The multistage fuzzy decision strategy is tested against heuristic-based mixed strategies using time-dependent and behaviour-dependent tactics from [16], which are able to adapt to the opponent's behaviour. The results shown in Figure 2 demonstrate that the strategy using the multistage fuzzy decision approach obtains on average higher utilities than the heuristic-based negotiation strategies (the red/dark difference on top showing the gain in utility when using the proposed strategy). However, it should be noted that the gain in utility depends to a high degree on the choice of the reference cases as they constitute the course of actions that can possibly lead to an agreement. The main advantage is that the agents can adjust or add new negotiation patterns and constraints in future negotiations in order to increase their utility gain and the number of agreements.

V. Stage in the Research Project

The PhD research project is in the stage where negotiation strategies have been modelled and evaluated with the multistage fuzzy decision approach with finite termination times using fuzzy state transitions. Currently, models with fuzzy and implicitly specified termination times are applied to the negotiation process and the generation of strategies. Furthermore, different approaches for modelling the state and action space are considered such as using concessions instead of offers. The evaluation of the above mentioned variants of the model are ongoing where, first, more than two cases are used and, second, the approach is compared with other experience based methods, such as case-based reasoning which also allow to use only limited information (e.g. two past interactions). Finally, the mechanism will be transformed to and tested in multi-issue negotiation scenarios.

VI. Conclusion

In this paper, the approach of multistage fuzzy decision making has been presented and how it is applied to model adaptive negotiation strategies using only limited and uncertain information about the opponent’s behaviour and the environment. It could be shown that the proposed model is able to adapt to different opponent’s behaviour based on the uncertain knowledge, which was derived from only two reference cases.

REFERENCES


