THE IEEE
Computational Intelligence
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Computational Intelligence

June 2003 Vol. 2 No. 1 (ISSN 1727-5997)

On-line version: http://www.comp.hkbu.edu.hk/~cib (ISSN 1727-6004)

The IEEE Computational Intelligence Bulletin

Aims and Scope

The IEEE Computational Intelligence Bulletin is the official publication of the Technical Committee on Computational Intelligence (TCCI) of the IEEE Computer Society, which is published twice a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCCI Executive Committee
2) Feature Articles
3) R & D Profiles (R & D organizations, interview profiles on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCCI sponsored or important/related activities)

Materials suitable for publication at the IEEE Computational Intelligence Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Computational Intelligence*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. Jiming Liu; Email: jiming@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997 (printed) 1727-6004 (on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing — Google (www.google.com), The ResearchIndex (citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and DBLP Computer Science Bibliography (www.informatik.uni-trier.de/∼ley/db/index.html).*

USC/ISI Polymorphic Robotics Laboratory

SELF-CONFIGURABLE AND ADAPTABLE ROBOTS

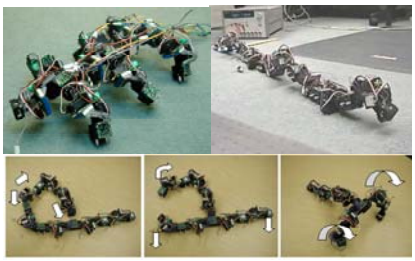
I. THE POLYMORPHIC ROBOTICS LAB

Over the last 30 years, the University of Southern California's Information Sciences Institute (ISI) has emerged as one of the world's leading research centers in the fields of computer science and information technology. It would be impossible to cover the breadth of their research in a short article like this, so we focus on a very small laboratory that is making big news in the robotics world.

The ISI Polymorphic Robotics Laboratory (PRL) is one of six laboratories associated with the USC School of Engineering's Center for Robotics and Embedded Systems. Led by Wei-Min Shen and Peter Will, this laboratory is a leader in self-reconfigurable and adaptive robot research, with projects including self-assembly for Space Solar Power Systems (SOLAR), the self-reconfigurable robot CONRO, the robot soccer team DREAMTEAM, the Intelligent Motion Surface, and the indoor navigation robot YODA.

PRL researchers develop modular hardware systems and teams of independently-controlled autonomous agents that work in synergy. Their robot soccer team was ranked #1 in the 1997 RoboCup contest due to their speed and quick reactions. The lab is currently organizing a new team.

In this article we highlight two related projects that are particularly good examples of some of the novel ideas being generated by this laboratory.



CONRO units can combine in many different configurations

USC
INFORMATION
SCIENCES
INSTITUTE

II. MODULAR ROBOTS GROW BY MELDING

Real-life "Transformer" robots can automatically reconfigure as needed for different applications.

PRL researchers have designed robotic units that can knit themselves together in a variety of ways, automatically taking on different behaviors depending on their positions. Identical modular units can autonomously look for and find each other, link themselves to one another and then, when united, work effectively as a single unified whole.

Metamorphing robots − known as CONROs, for CONfigurable RObots − may soon be worming their way through small crevices in earthquake debris or at a fire scene. Once inside, the tail of the worm could disconnect and reattach as legs. Each CONRO unit could be equipped with a different sensor, speaker, or even be able to deliver doses of medicine. Microscopic versions might work their way into keyholes and unlock doors, or even assemble inside the body to form surgical instruments.

The prototype CONRO units are about three inches long, consisting of a few small electric motors, a processing unit, and an active end that can move back and forth and up and down. This end has plugs on three sides that fit into receptacles on the front ends of other devices.

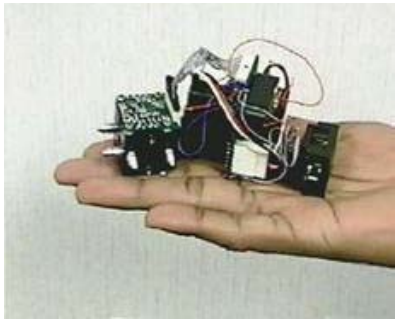
Separated units communicate using infrared signals, maneuvering their coupling units into a lock in cooperative, coordinated fashion. They use very simple messages called "hormones" by analogy to biological chemical messengers. In much the same way, CONRO's hormones are passed between nearest neighbors so their main data is type and locality.

Shen said that such software allows "bifurcation, unification, and behavior shifting" by the modules. The units can unite themselves into larger wholes, or divide themselves up into smaller ones. "If a six-unit snake splits in half," explained Shen, "you get two smaller, three-unit snakes that function as the larger one did."

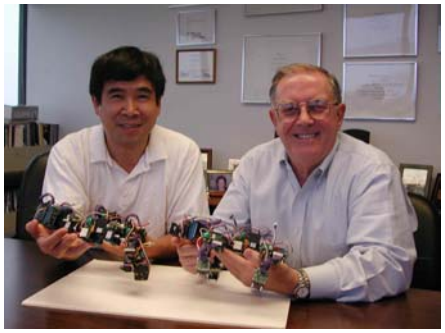
"Behavior shifting" means that identical individual units can exhibit different behavior according to their position in the assembly.

More than a year ago, the working group succeeded in the seemingly modest step of having a snake of six robots find and link to its own tail to form a ring. In another configuration, with six units linked as legs to a seventh and eighth units which serve as a body, an insect-like creature emerges which can walk on six legs, moving three at a time.

A recent development is a snake that reconstructs itself into a T-shaped assembly that "flies" or "swims" by moving the sides of the T crossbar like wings or fins.



One CONRO unit.



Wei-Min Shen, left, and Peter Will, holding hormone software controlled CONRO modules. They recently received the Phi Kappa Phi Faculty Recognition award for this work.

Peter Will, who in 1990 received the robotics profession's highest honor, the International Engelberger Prize, directs the CONRO project. He notes that creation of truly capable metamorphing robots will require improvements in many areas, including the development of chips that do more with less power.

ISI's Polymorphic Robots Laboratory is not the only site doing research on modular, self-assembling robots. Xerox Parc has long been working on a parallel project, called PolyBots.

CONRO's reconfiguration behaviors are autonomous and controlled in a totally distributed manner, Shen says, which distinguishes CONRO from many existing chain-typed self-reconfigurable robots like those from Xerox Parc.

The demands on software to achieve this performance are daunting. "The robots must recognize the conditions that dictate a change in form, must determine the proper new form to assume, and be able to do so quickly and efficiently under confused, real world conditions." he said. "These are major challenges. Nevertheless, the rewards for successful implementation of this technology make a vigorous effort worthwhile, and we are cheered by the successes we have so far achieved."

III. Robot Space Cowboys

A unique University of Southern California design for self-organizing robots controlled by "hormonal" software is moving toward space.

At the Robosphere 2002 conference held at the NASA Ames Research Center in Silicon Valley November 14-15, Wei-Min Shen presented an overview of an audacious project to have pieces of the proposed half-mile-long Space Solar Power System satellite assemble themselves without the help of astronauts.

Shen and Will's new SOLAR space station proposal, funded by a consortium including NASA, the NSF, and the Electric Power Research Institute (EPRI), proposes to use the CONRO architecture on a gigantic scale.

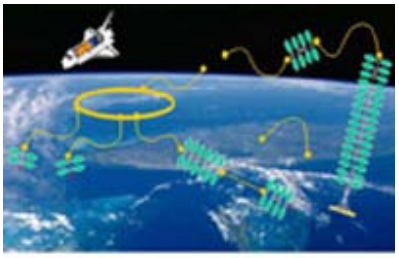
They propose a self-assembling space station consisting of two species of robotic devices, both controlled by the same software.

One species will be the parts that actually make up the station: solar power units, including necessary utility conduits. Each of these will have a microprocessor running hormonal software. Sets of contiguous units will, once released into space, arrange themselves into the desired configuration.

When these subassemblies are ready, they will signal and alert a second species of robot, the "free-flying intelligent fiber rope matchmaker units," or *whips*.

Whips consist of two modular robot units connected by a long connector line that can shorten or lengthen at the direction of the software. They will also have solar-powered rockets, enabling them to move in space, GPS sensors to find their position, communicators, and connectors.

When a completed subassembly signals, a whip will maneuver toward it, lock on, and wait for a call from a second assembly. The free end of the whip will fly to the second assembly and lock on to it. Then the whip pulls the parts together by shortening the connector line. Once the two parts are mated, the whip unit can fly off to find other parts to assemble. The design, said Will, combines the advantages of free-flying and tethered systems.



Schematic diagram of architecture of self-assembling solar power satellite. Seeker "whip" units, powered at both ends, listen for signals from subassemblies, find them, and pull them together.

In the laboratory, Shen and Will have modeled the concept in two-dimensional form, working with an air-hockey table. The prototype whips find parts by sensing their infrared signals, maneuver next to them using built-in fans, mechanically lock on, and pull units together using a motorized cable.

"This will give both the hardware and software a realistic test," said Shen. Researcher Harshit Suri has built a first prototype unit.

Shen, Will, and ISI collaborator Behnam Salemi published a detailed paper, "Hormone-Inspired Adaptive Communi-
cation and Distributed Control for CONRO Self-Reconfigurable Robots," in *IEEE Transactions on Robotics and Automation* on October, 2002. They have recently applied for a U.S. patent on the technology.

Working with Shen and Will in the field of space assembly are two faculty members from the USC School of Engineering: Berokh Khoshnevis of the department of industrial and systems engineering; and George Bekey, of the department of computer science. Along with Suri and Salemi, Yusuf Akteskan is working on the space system project.

Contact Information

Director: Wei-Min Shen
E-mail: shen@isi.edu
Phone: (310) 448-8710
Polymorphic Robotics Laboratory
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
Website: http://www.isi.edu/robots/

THE 18^{th} ACM SYMPOSIUM ON APPLIED COMPUTING

By Ronaldo Menezes, Vice-Chair of ACM SAC 2003

The ACM symposium on applied computing (SAC) is organized yearly having the Special Interest Group on Applied Computing (SIGAPP) as the sole sponsor. Since its 1999 edition, the symposium has been alternated between USA and Europe, every year attracting a large number of submissions and attendees.

The 18^{th} edition of the symposium was held in Melbourne, Florida from the 9^{th} to the 12^{th} of March, hosted by the Department of Computer Sciences at the Florida Institute of Technology.

This year's officials were: Gary Lamont (Symposium Chair), Ronaldo Menezes (Symposium Vice-Chair), George Papadopoulos (Program Chair), Hisham Haddad (Program Chair and Treasurer), Brajendra Panda (Publication Chair), Jan Carroll (Director), Warren Jones (Bioinformatics Director), William Shoaff (Local Arrangements Chair) and Ryan Stansifer (Tutorials Chair).

I. TUTORIALS

In its current format, tutorial sessions are held on Sunday while the technical sessions are Monday through Wednesday. The call four tutorials attracted 18 high level proposals from which only 4 could be selected. The format chosen was to offer 4 half-day tutorials. Given the proposals submitted and the interest of the local industry, the symposium accepted two tutorials in Wireless Networks, forming a full day theme on Wireless. The other two tutorials were on Semantic Web and on Complex Event Processing. Tutorials took place at the campus of the Florida Institute of Technology.

a) Wireless/Mobile Network Security by Dr. S. R. Subramanya from the University of Missouri-Rolla: The tutorial gave an overview of: (i) the principles and practices of various aspects of traditional network security, (ii) the issues in mobile and wireless security, and (iii) techniques and applications of mobile and wireless security.

b) Resource and Mobility Management in Next Generation Wireless Systems by Dr. Sajal K. Das from the University of Texas at Arlington: The tutorial aimed at bringing out the research and technological challenges in the resource and mobility management in next generation heterogeneous wireless systems. It also provided some emerging solutions for this problem to support wireless data networking.

c) Semantic Web and Ontologies by Raphael Volz from the University of Karlsruhe: In this tutorial, the author motivated the building and driving of ontology-based Semantic Web applications. He described Semantic Web standards as well as mechanisms to represent, engineer and use ontologies in Semantic Web applications. The tutorial was divided into four main parts: technological foundations, ontology representation, ontology engineering, and Semantic Web applications. This was a very popular tutorial attracting the interest of the majority of the symposium registered tutorial attendees.

d) Complex Event Processing in Distributed Enterprise Systems by Prof. David Luckham from Stanford University: Prof. Luckham has held faculty and invited faculty positions in mathematics, computer science and electrical engineering at eight major universities in Europe and the United States. He was one of the founders of Rational Software Inc. in 1981. His tutorial covered the basic concepts of Complex Event Programming: (i) the need for new technologies to manage the electronic enterprise, (ii) basic events and complex events, (iii) relationships of time, causality, independence and aggregation between events, (iv) hierarchical structure in enterprises and how to precisely define corresponding complex event hierarchies, (v) event pattern languages and rules, (vi) applying event hierarchies to enterprise management.

II. KEYNOTE ADDRESSES

One of the strengths of SAC is the ability to attract applied researchers from diverse areas. This year's keynote addresses is a clear example of this diversity. Following the successful format used in early editions of the symposium, SAC'03 included 2 keynote addresses and 1 luncheon speaker. On Monday, Dr. Lawrence Hunter presented his keynote address entitled *The Era of Biognostic Machinery*. In his talk he argued that knowledge-based approaches, ranging from graphical statistical models with informative priors, to rule-based inference and knowledge-based information extraction from natural language are the best way to meet the challenges faced by the need of analyzing and interpreting increasing amounts of data generated by molecular instrumentation. Dr. Hunger was invited by the former Special Interest Group on Biomedical Computing (SIGBIO). SIGBIO used to be a co-sponsor of SAC and in the

last couple of years have contributed to the symposium by sponsoring a keynote speaker.

Another very interesting keynote, that helped to solidify SAC's relation to industry, was given by Dr. Richard Simonian, Vice President of Engineering for the Government Communication Systems Division of Harris Corporation in Melbourne, Florida. In his talk he tackled the important and controversial problem of bridging research and industry, and the meaning of the software engineering profession. He discussed how fundamental system architecture principles are helpful to drive meaningful research and development and how to apply processes for development both "in the small" and "in the large".

The luncheon speaker was Dr. Lee Weng, Director of Applied Research of Rosetta Biosoftware, a division of Rosetta Inpharmatics LLC. In his talk, he discussed the challenges the industry is facing in providing powerful analysis and data management tools that can be used to revolutionize the field of computational biology.

III. The Technical Sessions

The symposium is divided into technical tracks. This year's edition counted with 21 tracks carefully selected from 30 track proposals. The subjects ranged from Applications to Healthcare to Computer Security, from Agents to Web Applications (see Fig. 1). Across all the tracks, the symposium received 525 papers from which only 194 were published. The Data Mining track alone (the largest track in the symposium) received an excess of 60 papers submitted. These numbers reflect a 37% acceptance rate making SAC 2003 not only the most successful edition of SAC so far, but also one of the most popular and competitive conferences in the international field of applied computing

Besides the growth in the number of submissions, we can notice that there are a number of authors that have presented their work at SAC for several years. This demonstrates a singular aspect of the symposium, that presenters enjoy the experience and organization of the symposium and are eager to attend in future years.

IV. Next Edition

In 2004, SAC will take place in the beautiful island of Cyprus. Following the success of the last few years, we believe SAC will continue to grow and continue to attract high quality works from all over the world. The call for papers for 2004 as well as the call for track proposals are already available.

A few new people will be joining the organization of the SAC which warrants the input of fresh new ideas.

V. Contact Information

More information about SAC 2004 can be found at `www.acm.org/conferences/sac/sac2004/`

For information about the Department of Computer Sciences at the Florida Institute of Technology, go to `cs.fit.edu`.

For information about the University of Cyprus, go to `www.ucy.ac.cy`.

Dr. Ronaldo Menezes is an Assistant Professor in Computer Sciences at Florida Tech, USA. His research interests are Coordination Systems and Natural-forming multi-agent Systems (aka Swarms) (`rmenezes@cs.fit.edu`).

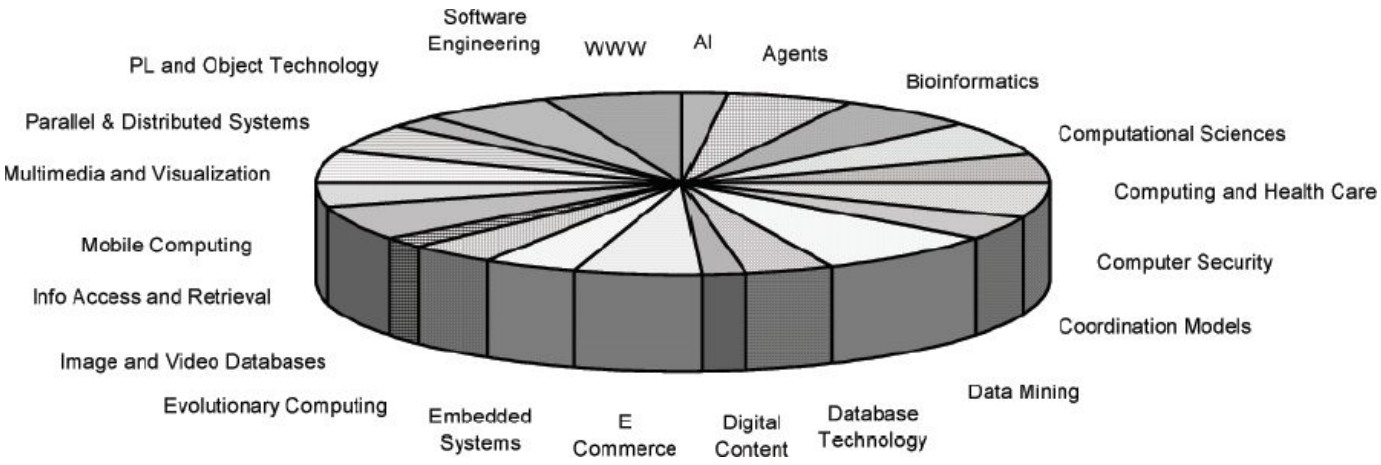


Fig. 1. Subjects of papers accepted by SAC 2003

Multi-Database Mining[1]

Shichao Zhang[2,3], Xindong Wu[4] and Chengqi Zhang[2]

***Abstract*— Multi-database mining is an important research area because (1) there is an urgent need for analyzing data in different sources, (2) there are essential differences between mono- and multi-database mining, and (3) there are limitations in existing multi-database mining efforts. This paper designs a new multi-database mining process. Some research issues involving mining multi-databases, including database clustering and local pattern analysis, are discussed.**

I. Introduction

THE increasing use of multi-database technology, such as computer communication networks and distributed, federated and homogeneous multi-database systems, has led to the development of many multi-database systems for real-world applications. For decision-making, large organizations need to mine the multiple databases distributed throughout their branches. In particular, as the Web is rapidly becoming an information flood, individuals and organizations can take into account low-cost information and knowledge on the Internet when making decisions. The data of a company is referred to as internal data whereas the data collected from the Internet is referred to as external data. Although external data assists in improving the quality of decisions, it generates a significant challenge: how to efficiently identify quality knowledge from multi-databases [26], [30], [31]. Therefore, large companies may have to confront the multiple data-source problem. Recently, the authors have developed local pattern analysis, a new multi-database mining strategy for discovering some types of potentially useful patterns that cannot be mined with traditional data mining techniques. Local pattern analysis discovers high-performance patterns from multi-databases.

There are two fundamental problems that prevent local pattern analysis from widespread applications. First, the data collected from the Internet is of poor quality that can disguise potentially useful patterns. For example, a stock investor might need to collect information from outside data sources when making an investment decision. If fraudulent information collected on the Internet is directly applied to investment decisions, the investor might lose money. In particular, much work has been built on consistent data. With distributed data mining algorithms it is assumed that the databases do not conflict with each other. However, reality is much more inconsistent, and inconsistency must be resolved before a mining algorithm can be applied. These observations generate a crucial requirement: data preparation.

The second fundamental problem is efficient algorithms for identifying potentially useful patterns in multi-databases. Over the years, there has been a lot of work in distributed data mining. However, traditional multi-database mining still utilizes mono-database mining techniques. That is, all the data from relevant data sources is pooled to amass a huge dataset for discovery. This can destroy useful patterns. For example, a pattern like "80% of the 15 supermarket branches reported that their sales increased 9% when bread and milk were frequently purchased" can often assist in decision-making at a central company level. However, mono-database mining techniques may miss such a pattern in the centralized database. On the other hand, using our local pattern analysis, there can be huge amounts of local patterns. These observations generate a strong requirement for the development of efficient algorithms for identifying useful patterns in multi-databases.

There are other essential differences between mono- and multi-database mining. Both data and patterns in multi-databases present more challenges than those in mono-databases. For example, unlike in mono-databases, data items in multi-databases may have different names, formats and structures in different databases. They may also conflict with each other.

In this paper we present a multi-database mining system through defining a new process for multi-database mining. The rest of this paper is organized as follows. Section II illustrates the role of multi-database mining in real-world applications. Section III describes multi-database mining problems. Section IV analyzes the differences between mono- and multi-database mining by demonstrating the features of data in mono- and multi-databases. Section V recalls the research into multi-database mining. Section VI designs a process for multi-database mining. Section VII discusses the features of our proposed multi-database mining.

II. Multi-Database Mining in Real-World Applications

Business, government and academic sectors have all implemented measures to computerize all, or part of, their daily functions [12]. An interstate (or international) company consists of multiple branches. The National Bank of Australia, for example, has many branches in different locations. Each branch has its own database, and the bank data is widely distributed and thus becomes a multi-database problem (see Fig. 2).

In Fig. 2, the top level is an interstate company (IC). This IC is responsible for the development and decision-making for

[1]This research has been partially supported by the Australian Research Council under grant number DP0343109, the Guangxi Natural Science Funds, and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number DAAD19-02-1-0178.

[2]Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, Australia {zhangsc,chengqi}@it.uts.edu.au

[3]State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, China

[4]Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA xwu@cs.uvm.edu

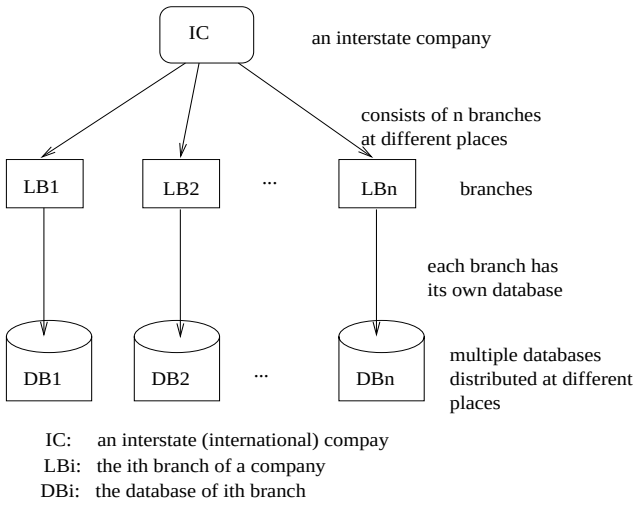


Fig. 2. An interstate company and its branches

the entire company. The middle level consists of n branches LB_1, LB_2, $\cdots$, LB_n. The bottom level consists of n local databases DB_1, DB_2, $\cdots$, DB_n of the n branches.

Fig. 2 illustrates the structure of a two-level interstate company. In the real world, the structure of an interstate company is usually more complicated, and each branch may also have multi-level sub-branches.

Many organizations have a pressing need to manipulate all the data from their different branches rapidly and reliably. This need is very difficult to satisfy when the data is stored in many independent databases, and the data is all of importance to an organization. Formulating and implementing queries requires data from more than one database. It requires knowledge of where all the data is stored, mastery of all the necessary interfaces and the ability to correctly combine partial results from individual queries into a single result.

To respond to these demands, researchers and practitioners have intensified efforts on developing appropriate techniques for utilizing and managing multi-database systems. Hence, developing multi-database systems has become an important research area.

Also, the computing environment is becoming increasingly widespread through the use of Internet and other computer communication networks. In this environment, it has become more critical to develop methods for building multi-database systems that combine relevant data from many sources and present the data in a form that is comprehensible for users, and provide tools that facilitate the efficient development and maintenance of information systems in a highly dynamic and distributed environment. One important technique within this environment is the development of multi-database systems. This includes managing and querying data from the collections of heterogeneous databases.

While multi-database technology can support many multi-database applications, it would be useful and necessary to mine these multi-databases to enable efficient utilization of the data. Thus, the development of multi-database mining is both a challenging and critical task.

Some essential differences between mono- and multi-database mining will be demonstrated below. We will show that traditional multi-database mining techniques are inadequate for two-level applications within large organizations such as interstate companies.

III. Multi-Database Mining Problems

An interstate company often consists of multi-level branches. Without loss of generality, this paper simplifies each interstate company as a two-level organization (a central company and multiple branches), as depicted in Fig. 2. Each branch has a database and the database is simplified as a relation or a table for our mining purposes.

Fig. 2 can be used to demonstrate that there are fundamental differences between mono- and multi-database mining. For example, multi-database mining may be restricted by requirements imposed by two-level decisions: the central company's decisions (global applications) and branch decisions (local applications). For global applications and for corporate profitability, central company headquarters are more interested in patterns (rather than the original raw data) that have the support of most of its branches, and those patterns are referred to as high-vote patterns hereafter. In local applications, a branch manager needs to analyze the data to make local decisions.

Two-level applications in an interstate company are depicted in Fig. 3.

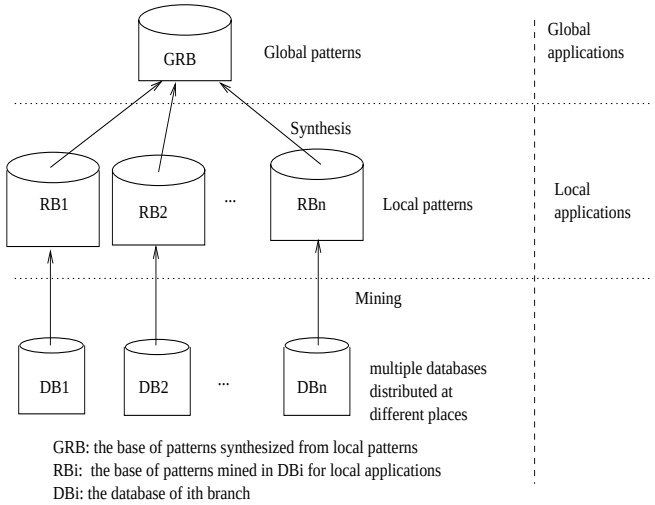


Fig. 3. Two level applications in an interstate company

In Fig. 3, the bottom level consists of n local databases DB_1, DB_2, $\cdots$, DB_n of n branches within an interstate company. The middle level consists of n sets RB_1, RB_2, $\cdots$, RB_n of local patterns discovered from databases DB_1, DB_2, $\cdots$, DB_n, respectively. These local patterns can be used for decision-making within branches (local applications). The top level is a set of global patterns that are synthesized from the n sets RB_1, RB_2, $\cdots$, RB_n. These global patterns are used for the overall company's decision-making (global applications).

One possible way for multi-database mining is to integrate all the data from these databases to amass a huge dataset for discovery by mono-database mining techniques. However, there are important challenges and difficulties involved in applying this method to real-world applications. We will discuss these challenges and difficulties in detail in Section V-B.

In Fig. 3 each database has been mined at each branch for use in local applications. While collecting all data together from different branches might produce a huge database and lose some important patterns for the propose of centralized processing, forwarding the local patterns (rather than the original raw data) to central company headquarters provides a feasible means of dealing with multiple database problems. The patterns forwarded from branches are called *local patterns*.

However, the number of forwarded local patterns may be so large that browsing the pattern set and finding interesting patterns can be rather difficult for central company headquarters. Therefore, it can be difficult to identify which of the forwarded patterns (including different and identical ones) are really useful at the central company level.

IV. Differences Between Mono- And Multi-Database Mining

The previous sections have indicated that there are essential differences between mono- and multi-database mining. This section illustrates these differences using the features of data and patterns in mono- and multi-databases.

A. Features of Data in Multi-databases

There are many ways to model a given real-world object (and its relationships with other objects) in, for example, an interstate company, depending on how the model will be used [12]. Because local databases are developed independently with differing local requirements, a multi-database system is likely to have many different models, or representations, for similar objects. Formally, a multi-database system is a federation of autonomous, and possibly heterogeneous, database systems used to support global applications and concurrent accesses to data stored in multiple databases [12].

We now illustrate data features in multi-databases.

1) *Name differences*. Local databases may have different conventions for the naming of objects, leading to problems with synonyms and homonyms.

A synonym means that the same data item has a different name in different databases. The global system must recognize the semantic equivalence of the items and map the differing local names to a single global name. A homonym means that different data items have the same name in different databases. The global system must recognize the semantic difference between items and map the common names to different global names.

2) *Format differences*. Many analysis or visualization tools require that data be in particular formats within branches. Format differences include differences in data type, domain, scale, precision, and item combinations.

An example is when a part number is defined as an integer in one database and as an alpha-numeric string in another. Sometimes data items are broken into separate components in one database while the combination is recorded as a single quantity in another.

Multi-database systems typically resolve format differences by defining transformation functions between local and global representations. Some functions may consist of simple numeric calculations such as converting square feet to acres. Others may require tables of conversion values or algorithmic transformations. A problem in this area is that the local-to-global transformation (required if updates are supported) may be very complex.

3) *Structural differences*. Depending on how an object is used in a database system, it may be structured differently in different local databases.

A data item may have a single value in one database and multiple values in another. An object may be represented as a single relation in one location or as multiple relations in another. The same item may be a data value in one location, an attribute in another, and a relation in a third. So the data often has discrepancies in structure and content that must be cleaned.

4) *Conflicting data*. Databases that model the same real-world object may have conflicts within the actual data values recorded.

One system may lack some information due to incomplete updates, system errors, or insufficient demands to maintain such data. A more serious problem arises when two databases record the same data item but assign it different values. The values may differ because of an error, or because of valid differences in the underlying semantics.

5) *Distributed data*. In most organizations, data is stored in various formats, in various storage media, and with various computers.

Therefore, data is created, retrieved, updated and deleted using various access mechanisms.

6) *Data sharing*. A major advantage of multi-database systems is the means by which branch data and sources can be shared.

In an interstate company, each of its branches has individual functions, data and sources. These branches can interact and share their data when they cannot solve problems that are beyond their individual capabilities.

7) *Data for two-level applications*. Comprehensive organizations have two-level decisions: central company's decisions (global applications) and branch decisions (local applications).

The above features demonstrate that data in multi-databases is very different from data in mono-databases.

B. Features of Patterns in Multi-databases

Generally, patterns in multi-databases can be divided into (1) local patterns, (2) high-vote patterns, (3) exceptional patterns, and (4) suggested patterns.

1) *Local patterns.* In an interstate company, local branches need to consider the original raw data in their databases so they can identify local patterns for local decisions.

Each branch of an interstate company has certain individual functions. The branch must design its own plan and policy for development and competition. It therefore needs to analyze data only in their local databases to identify local patterns. Each branch can then share these patterns with other branches. More importantly, they can forward their local patterns to the central company when global decisions need to be made.

2) *High-vote patterns.* These are patterns that are supported/voted for by most branches. They reflect common characteristics among branches and are generally used to make global decisions.

When an interstate company makes a global decision, the central company headquarters are usually interested in local patterns rather than original raw data. Using local patterns, they can learn what is supported by their branches. High-vote patterns are helpful in making decisions for the central company.

3) *Exceptional patterns.* These are patterns that are strongly supported/voted for by only a few branches. They reflect the individuality of branches and are generally used to create special policies specifically for those branches.

Although high-vote patterns are useful in reaching decisions for an interstate company, the headquarters are also interested in viewing the exceptional patterns used for making special decisions at only a few of the branches. Exceptional patterns may also be useful in predicting/testing the sales of new products.

4) *Suggested patterns*. These are patterns that have less votes than the minimal vote (written as $minvote$) but are very close to $minvote$.

The minimal vote is given by the user or a domain expert. It means that if a local pattern has votes equal to, or greater than, $minvote$, the local pattern becomes a global pattern, and is known as a high-vote pattern. Under the threshold $minvote$, there may be some local patterns that have less votes than $minvote$ but are very close to it. We call these patterns suggested patterns and they are sometimes useful in global decisions.

It is important to note that *local patterns also inherit the features of data in multi-databases*.

The above differences in data and patterns in multi-database systems demonstrate that multi-database mining differs from mono-database mining. This invites the exploration of efficient mining techniques for identifying novel patterns in multi-databases such that patterns can serve two-level applications in large organizations.

V. Related Work

A. Existing Research Efforts on Multi-Database Mining

If an interstate company is a comprehensive organization where its databases belong to different types of businesses and have different meta-data structures, the databases would have to be classified before the data is mined. For example, if a company like Coles-Myer has 25 branches including 5 supermarkets for food, 7 supermarkets for clothing, and 13 supermarkets for general commodities, these databases would first have to be classified into three clusters according to their business types before they are mined. Therefore, a key problem in multi-database mining is how to effectively classify multi-databases.

To mine multi-databases, the first method (mono-database mining technique) is to put all the data together from multiple databases to create a huge mono-dataset. There are various problems with this approach and we will discuss them in Section V-B.

In order to confront the size of datasets, Liu, Lu and Yao have proposed an alternative multi-database mining technique that selects relevant databases and searches only the set of all relevant databases [15]. Their work has focused on the first step in multi-database mining, which is the identification of databases that are most relevant to an application. A relevance measure was thus proposed to identify relevant databases for mining with an objective to find patterns or regularity within certain attributes. This can overcome the drawbacks that are the result of forcedly joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application.

Identifying relevant databases in [15] is referred to as database selection. In real-world applications, database selection needs, however, multiple times to identify relevant databases to meet different applications. In particular, the users may need to mine their multi-databases without specifying any application, and in this case, the database selection approach does not work. The database selection approach is application-dependent.

While data mining techniques have been successfully used in many diverse applications, multi-database mining has only been recently recognized as an important research topic in the data mining community. Yao and Liu have proposed a means of searching for interesting knowledge in multiple databases according to a user query. The process involves selecting all interesting information from many databases by retrieval. Mining only works on the selected data [28].

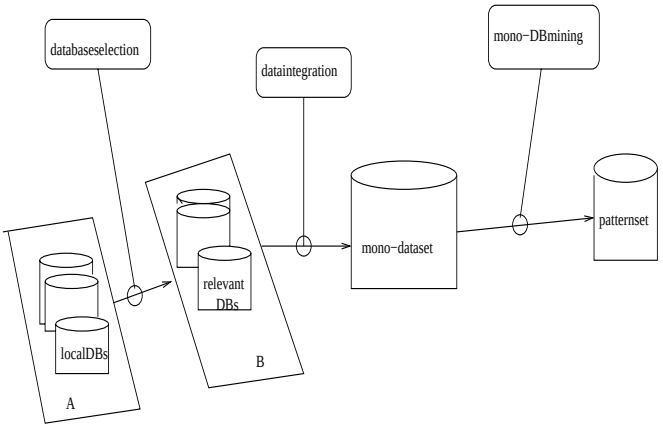


Fig. 4. The traditional process of multi-database mining

Based on [15], [28], Fig. 4 illustrates the functions used in existing multi-database mining. We call this process the traditional process. Area 'A' contains n sets of local databases

in an interstate company, where 'localDBs' stand for a set of local databases. 'databaseselection' is a procedure of the application-dependent database classification that identifies databases most relevant to an application. Area 'B' contains all databases that are relevant to an application. 'dataintegration' is a procedure that integrates all data in the relevant databases into a single dataset, called a 'mono-dataset'. Meanwhile, 'mono-DBmining' is a procedure that uses mono-database mining techniques to mine the integrated mono-dataset. 'patternset' is a set of the discovered patterns in the mono-dataset integration.

Zhong *et al.* have proposed a way of mining peculiarity patterns from multiple statistical and transaction databases based on previous work [31]. A peculiarity pattern is discovered from the peculiar data by searching the relevance among the peculiar data. Roughly speaking, a data item is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like an exception pattern from the viewpoint of describing a relatively small number of objects, the peculiarity pattern represents a well-known fact with common sense, which is a feature of the general pattern.

A related research effort is distributed data mining (DDM) that deals with different possibilities of data distribution. A famous effort is hierarchical meta-learning [18] which has a similar goal of efficiently processing large amounts of data. Meta-learning starts with a distributed database or a set of data subsets of an original database, concurrently runs a learning algorithm (or different learning algorithms) on each of the subsets, and combines the predictions from classifiers learned from these subsets by recursively learning 'combiner' and 'arbiter' models in a bottom-up tree manner [18]. The focus of meta-learning is to combine the predictions of learned models from the partitioned data subsets in a parallel and distributed environment.

Other related research projects are now briefly reviewed. Wu and Zhang have advocated an approach for identifying patterns in multi-database by weighting [26]. Ribeiro, Kaufman and Kerschberg have described a way of extending the INLEN system for multi-database mining by incorporating primary and foreign keys, as well as developing and processing knowledge segments [20]. Wrobel has extended the concept of foreign keys to include foreign links, since multi-database mining also involves accessing non-key attributes [25]. Aronis *et al.* introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases spread across the network [4]. Kargupta *et al.* have built a collective mining technique for distributed data [14], [13]. Grossman *et al.* have built a system, known as Papyrus, for distributed data mining [9], [22]. Existing parallel mining techniques can also be used to deal with multi-databases [5], [7], [18], [19], [21].

The above efforts have provided good insights into multi-database mining. However, they are inadequate for identifying two new types of patterns: high-vote patterns and exceptional patterns, which reflect the distributions of local patterns.

B. Limitations of Mono-Database Mining for Dealing with Multiple Databases

Despite there being several methods of multi-database mining, most of them are still closely modeled on techniques for mono-database mining. This leads to a number of serious concerns and problems.

1) Due to the difficulty of data preparation, most work on multi-database mining has been built on quality data, and it is assumed that the data in different data sources is nicely distributed and contains consistent and correct values. However, existing data preparation focuses on single databases [29]. Because there are essential differences between multi- and mono-databases, there is a significant need of preparing the data in multi-databases.
2) Putting all the data from relevant databases into a single database can destroy some important information that reflects the distribution of patterns. These patterns may be more important than the patterns present in the single database in terms of global decision-making by a centralized company. Hence, existing techniques for multi-databases mining are inadequate.

We have provided an example in this regard in the introduction. In some cases, each branch of an interstate company, large or small, has equal power of voting for patterns involved in global decisions. For global applications, it is natural for the central company headquarters to be interested in the patterns voted for by most of the branches or exceptional patterns. It is therefore inadequate in multi-database mining to utilize existing techniques used for mono-databases mining.

3) Collecting all data from multi-databases can amass a huge database for centralized processing.

It may be an unrealistic proposition to collect data from different branches for centralized processing because of the huge data volume. For example, different branches of Wal-Mart receive 20 million transactions a day. This is more than the rate at which data can be feasibly collected and analyzed using today's computing power. The French Teletel system has 1500 separate databases [12].

Parallel mining is sometimes unnecessary as there are many techniques such as sampling and parallel algorithms, for dealing with large databases.

A better approach is to first classify the multiple databases. The data from a class of databases can then be put into a single database for knowledge discovery utilizing existing techniques.

4) Forwarding all rules mined in branches to a central company. The number of forwarded rules may be so large that browsing the rule set and finding interesting rules from it can be a difficult task. In particular, it is more difficult to identify which of the forwarded rules are genuinely useful.

One strategy may be to reuse all the promising rules discovered in branches because the local databases have been mined for local applications. However, to reuse the local rules and select from them, a method must be developed to (1) determine valid rules for the overall organization from the

amassed database, and (2) reduce the size of the candidate rules from multi-databases. The following problems arise: (a) any rule from a database has the potential to contribute in the construction of a valid rule for the overall organization, and (b) the number of promising rules from multi-databases can be very large before it is determined which ones are of interest.

5) Because of data privacy and related issues, it is possible that some databases of an organization may share their patterns but not their original databases.

Privacy is a very sensitive issue, and safeguarding its protection in a multi-database environment is of extreme importance. Most multi-database designers take privacy very seriously, and allow for some protection facilities. For resource sharing in real-world applications, sharing patterns is a feasible way. This is because (1) certain data, such as commercial data, is secret for competition reasons; (2) reanalyzing data is costly; and (3) inexperienced decision-makers don't know how to confront huge amounts of data. The branches of an interstate company must search their databases for local applications. Hence, forwarding the patterns (rather than the original raw data) to the centralized company headquarters presents a feasible way to deal with multi-database problems.

Even though all of the above limitations might not be applicable to some organizations, efficient techniques, such as sampling and parallel and distributed mining algorithms, are needed to deal with the amassed mono-databases. However, sampling models depend heavily on the transactions of a given database being randomly appended to the database in order to hold the binomial distribution. Consequently, mining association rules upon paralleling (MARP), which employ hardware technology such as parallel machines to implement concurrent data mining algorithms, are a popular choice [2], [5], [8], [16], [17], [21]. Existing MARP efforts endeavor to scale up data mining algorithms by changing existing sequential techniques into parallel versions. These algorithms are effective and efficient, and have played an important role in mining very large databases. However, in addition to the above five limitations, MARP has two more limitations when performing data mining with different data sources.

6) MARP does not make use of local rules at branches; nor does it generate these local rules. In real-world applications, these local rules are useful for the local data sources, and would need to be generated in the first instance.
7) Parallel data mining algorithms require more computing resources (such as massive parallel machines) and additional software to distribute components of parallel algorithms among different processors of parallel machines. Most importantly, it is not always possible to apply MARP to existing data mining algorithms. Some data mining algorithms are sequential in nature, and can not make use of parallel hardware.

From the above observations, it is clear that traditional multi-database mining is inadequate to serve two-level applications. This prompts the need to develop new techniques for multi-database mining.

VI. MDM: A New Process for Multi-Database Mining

As previously explained, there are three factors that illustrate the importance of multi-database mining: (1) there are many multi-databases already serving large organizations; (2) there are essential differences between mono- and multi-database mining; and (3) there are limitations in existing multi-database mining techniques. For these reasons, we have designed a high-performance prototype system for multi-database mining (MDM). Below we introduce our MDM design through defining a new process of multi-database mining and describing its functions.

A. Three Steps in MDM

There are various existing data mining algorithms that can be used to discover local patterns in local databases [1], [11], [23]. These include the paralleling algorithms mentioned above [18], [21]. Our MDM process focuses on local pattern analysis as follows.

> Given n databases within a large organization, MDM performs three steps: (i) searching for a good classification of these databases; (ii) identifying two types of new patterns from local patterns: high-vote patterns and exceptional patterns; and (iii) synthesizing local patterns by weighting.

The major technical challenge in MDM is to serve the two-level applications in large organizations, such as interstate companies. MDM is depicted in Fig. 5.

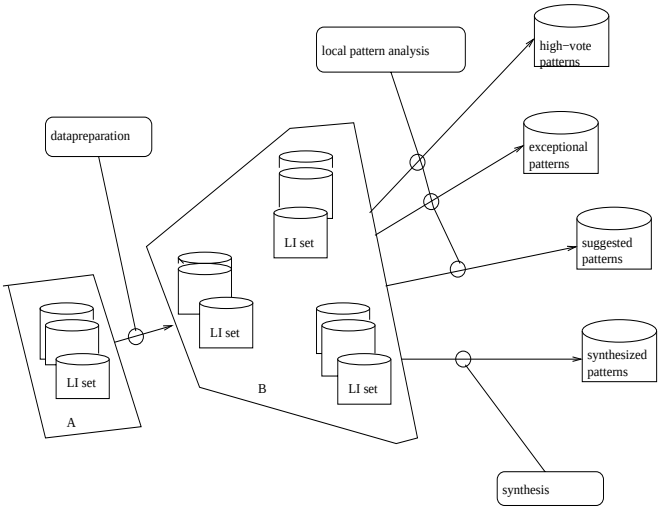


Fig. 5. The MDM process

In Fig. 5, area 'A' contains n sets of local patterns of an interstate company, where 'LIset' stands for a local pattern set; and 'datapreparation' is a procedure of application-independent database classification. After classifying the multi-databases, the local pattern sets are divided into several groups in area 'B'. For each group of local pattern sets, we use procedure 'localpatternanalysis' to search for patterns, such as high-vote

patterns, exceptional patterns, and suggested patterns. Procedure 'synthesis' is used to aggregate the local patterns in each group.

B. Research Issues in the MDM Process

In Fig. 5, three procedures, 'datapreparing', 'localpatternanalysis', and 'synthesis' are needed, as well as other procedures, to unify names of items and remove noise. Although the problem of unifying names of items and removing noise is also faced by multi-database systems [12], our MDM process focuses on issues raised from the three procedures in Fig. 5.

1) Data preparation can be more time consuming, and can present more challenges than mono-database mining. The importance of data preparation can be illustrated by the following observations: (1) real-world data is impure; (2) high-performance mining systems require quality data; and (3) quality data yields concentrative patterns. Therefore, the development of data preparation technologies and methodologies is both a challenging and critical task.

There are four key problems in data preparation: (i) developing techniques for cleaning data, (ii) constructing a logical system for identifying quality knowledge from different data sources, (iii) constructing a logical system for resolving inconsistency in different data sources, and (iv) designing application-independent database clustering.

(a) Developing techniques for cleaning data. Data cleaning techniques have been widely studied and applied in pattern recognition, machine learning, data mining and Web intelligence. For multi-database mining, distributed data cleaning presents more challenges than traditional data cleaning for single databases. For example, data may conflict within multi-databases. We need the following techniques to generate quality data for multi-database mining.

- Recover incomplete data: filling missing values, or expelling ambiguity;
- Purify data: consistency of data names and data formats, correcting errors, or removing outliers (unusual or exceptional values); and
- Resolve data conflicts: using domain knowledge or expert decisions to settle discrepancy.

(b) Constructing a logical system to identify quality knowledge from different data sources. As we argued previously, sharing knowledge (rather than the original raw data) presents a feasible way to deal with different data source problems [26]. Accordingly, assume that a data source is taken as a knowledge base[1]; a company (or a branch of the company) is viewed as a data source; and a rule has two possible values in a data source: true (if the data source supports the rule) or false (otherwise).

In the Web environment, the database from a company and information from different websites (called external data sources) can be treated as different data sources. External data sources may be subject to noise, and therefore, if a data source (a company or a branch) wants to form its own knowledge for data mining applications, the data source needs the ability of refining external knowledge. To do so, we advocate a logical system for identifying quality knowledge that focuses on the following epistemic properties.

- Veridicality. Knowledge is true.
- Introspection. A data source is aware of what it supports and of what it does not support.
- Consistency. A data source's knowledge is non-contradictory.

(c) Constructing a logical system for resolving inconsistency in different data sources. Traditional (positive) association rules can only identify companionate correlations among items. It is desirable in decision-making to catch the mutually-exclusive correlations among items that are referred to as negative associations. Therefore, we have developed a new method for identifying both positive and negative association rules in databases [27]. Negative association rules can increase the quality of decisions. However, in a multi-database environment, negative association rules can cause inconsistency within databases.

(d) Designing application-independent database clustering. To perform an effective application-independent database classification, we will have to (1) construct measurements for database relevance, (2) construct measurements of good classifications, and (3) design effective algorithms for application-independent database classification.

2) To provide effective multi-database mining strategies for identifying new patterns, we will develop four techniques for searching for new patterns from local patterns, that is, (a) design a local pattern analysis strategy; (b) identify high-vote patterns; (c) find exceptional patterns; and (d) synthesize local patterns by weighting.

(a) *Designing a local pattern analysis strategy*. Using traditional multi-database mining techniques, we can identify patterns, such as frequent itemsets, association patterns and classification patterns, by analyzing all the data in a database cluster. However, as mentioned in the introduction, these techniques can lose useful patterns. Therefore, analyzing local patterns is very important for mining novel and useful patterns in multi-databases.

On the other hand, for a large company, the number of local patterns may, however, be so large that browsing the pattern set and finding interesting patterns from it can be a difficult task for the company headquarters. In particular, it is harder to identify which of the local patterns are genuinely useful. Therefore, analyzing local patterns is also a difficult task.

In a multi-database environment, a pattern has attributes such as the name of the pattern, the rate voted for by branches, and supports (and confidences for a rule) in branches that vote for the pattern. In other words, a pattern is a super-point of the form

$$P(name, vote, vsupp, vconf).$$

In our system, we have designed a local pattern analysis strategy in [29] by using the techniques in [30]. The key problem to be solved is how to analyze the diverse projections of patterns in a multi-dimension space consisting of local patterns within a company.

[1]If a data source contains only data, we can transform it into knowledge by existing mining techniques.

(b) *Identifying high-vote patterns*. Within a company, each branch, large or small, has a power to vote for patterns for global decision-making. Some patterns can receive votes from most of the branches. These patterns are referred to as high-vote patterns. These patterns may be far more important in terms of global decision-making within the company.

Because traditional mining techniques cannot identify high-vote patterns, these patterns are regarded as novel patterns in multi-databases. In our system, we have designed a mining strategy for identifying high-vote patterns of interest based on a local pattern analysis. The key problem to be solved in this mining strategy is how to post-analyze high-vote patterns.

(c) *Finding exceptional patterns*. Like high-vote patterns, exceptional patterns are also regarded as novel patterns in multi-databases. But an exceptional pattern receives votes from only a few branches. While high-vote patterns are useful when a company is making global decisions, headquarters are also interested in viewing exceptional patterns when special decisions are made at only a few of the branches, perhaps for predicting the sales of a new product. Exceptional patterns can capture the individuality of branches. Therefore, these patterns are also very important.

(d) *Synthesizing patterns by weighting*. Although each branch has a power to vote for patterns for making global decisions, branches may be different in importance to their company. For example, if the sale of branch A is 4 times of that of branch B in a company, branch A is more important than branch B in the company. The decisions of the company can be reasonably partial to high-sale branches. Also, local patterns may have different supports in different branches. We will need a new strategy for synthesizing local patterns based on an efficient model for synthesizing patterns from local patterns by weighting [26].

VII. Features of the MDM Process

The MDM process in Section VI provides a new way for building multi-database mining systems. The main features of this process are as follows.

1) New mining techniques and methodologies can significantly increase the ability of multi-database mining systems.

Previous techniques in multi-databases mining were developed to search for patterns using existing mono-database mining. Although data in multi-databases can be merged into a single dataset, such merging can lead to many issues such as tremendous amounts of data, the destruction of data distributions, and the infiltration of uninteresting attributes. In particular, some concepts, such as regularity, causal relationships and patterns cannot be discovered if we simply search a single dataset, since the knowledge is essentially hidden within the multi-databases [31]. It is a difficult task to effectively exploit the potential ability of mining systems and it is one of the issues essential to achieve the objective of designing effective mining strategies.

Our multi-database mining strategy is to identify two types of patterns, high-vote patterns and exceptional patterns, from analyzing local patterns. Because previous techniques search patterns in the same way as in existing mono-database mining, they cannot discover high-vote patterns and exceptional patterns in multi-databases. Therefore, the high-vote and exceptional patterns are regarded as novel patterns. In particular, the discovery of these patterns can capture certain distributions of local patterns and assist global decision-making within a large company.

2) New mining techniques and methodologies can significantly improve the performance of multi-database mining systems.

As we argued previously, an interstate company must confront two-level decisions: the company's decisions (global applications) and the branches' decisions (local applications). For global applications, the company headquarters must tackle huge amounts of data and local patterns. Therefore, the development of high-performance systems for mining multi-databases is very important.

The local pattern analysis strategies can deliver two direct benefits: greatly reduce search costs by reusing local patterns, and offer more useful information for global applications.

For efficient multi-database mining, a key problem is how to analyze the data in the databases so that useful patterns can be found to support various applications. We have mentioned two new strategies in dealing with this difficult problem. The first strategy is to design an efficient and effective application-independent database classification. The second strategy is to develop a local pattern analysis for identifying novel and useful patterns.

VIII. Conclusion

As pointed out in [31], most of the KDD methods that have been developed are on the single universal relation level. Although theoretically, any multi-relational database can be transformed into a single universal relation, practically this can lead to many issues such as universal relations of unmanageable sizes, infiltration of uninteresting attributes, loss of useful relation names, unnecessary join operations, and inconvenience for distributed processing. In particular, some concepts, regularity, causal relationships, and rules cannot be discovered if we just search a single database since the knowledge hides in multiply databases basically.

This paper has shown that the problem of multi-database mining is challenging and pressing. In particular, due to essential differences between mono- and multi-databases, we have defined a new process of multi-database mining for our system.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, Database mining: A performance perspective. *IEEE Trans. Knowledge and Data Eng.*, Vol. 5, 6(1993): 914-925.

[2] R. Agrawal, J. Shafer: Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engneering*, 8(6) (1996): 962-969.

[3] J. Albert, Theoretical Foundations of Schema Restructuring in Heterogeneous Multidatabase Systems. In: *Proceedings of CIKM*, 2000: 461-470.

[4] J. Aronis *et al.*, The WoRLD: Knowledge discovery from multiple distributed databases. *Proceedings of 10th International Florida AI Research Symposium*, 1997: 337-341.

[5] J. Chattratichat, etc., Large scale data mining: challenges and responses. In: *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 1997: 143-146.

[6] P. Chan, An Extensible Meta-Learning Approach for Scalable and Accurate Inductive Learning. *PhD Dissertation*, Dept of Computer Science, Columbia University, New York, 1996.
[7] D. Cheung, J. Han, V. Ng and C. Wong, Maintenance of discovered association rules in large databases: an incremental updating technique. In: *Proceedings of International Conference on Data Engineering*, 1996: 106-114.
[8] D. Cheung, V. Ng, A. Fu and Y. Fu, Efficient Mining of Association Rules in Distributed Databases, *IEEE Trans. on Knowledge and Data Engg.*, 8(1996), 6: 911–922.
[9] R. Grossman, S. Bailey, A. Ramu, B. Malhi and A. Turinsky, The preliminary design of Papyrus: A system for high performance, distributed data mining over clusters. In: *Advances in Distributed and Parallel Knowledge Discovery*, AAAI Press/The MIT Press, 2000: 259-275.
[10] E. Han, G. Karypis and V. Kumar, Scalable Parallel Data Mining for association rules. In: *Proceedings of ACM SIGMOD*, 1997: 277-288.
[11] J. Han, J. Pei, and Y. Yin, Mining Frequent Patterns without Candidate Generation, *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data SIGMOD'00)*, Dallas, TX, May 2000.
[12] A. Hurson, M. Bright, and S. Pakzad, *Multidatabase systems: an advanced solution for global information sharing.* IEEE Computer Society Press, 1994.
[13] H. Kargupta, W. Huang, K. Sivakumar, and E. Johnson, Distributed clustering using collective principal component analysis. *Knowledge and Information Systems*, 3(4) (2001): 422-448.
[14] H. Kargupta, W. Huang, K. Sivakumar, B. Park, and S. Wang, Collective Principal Component Analysis from Distributed, Heterogeneous Data. In: Principles of Data Mining and Knowledge Discovery, 2000: 452-457.
[15] H. Liu, H. Lu, and J. Yao, Identifying Relevant Databases for Multidatabase Mining. In: *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1998: 210–221.
[16] J. Park, M. Chen, P. Yu: Efficient Parallel and Data Mining for Association Rules. In: *Proceedings of CIKM*, 1995: 31-36.
[17] S. Parthasarathy, M. J. Zaki, W. Li, Memory placement techniques for parallel association mining. *Proceedings of International Conference on Knowledge Discovery and Data Mining* 1998: 304-308.
[18] A. Prodromidis, S. Stolfo. Pruning meta-classifiers in a distributed data mining system. In: *Proc. of the First National Conference on New Information Technologies*, 1998: 151-160.
[19] A. Prodromidis, P. Chan, and S. Stolfo, Meta-learning in distributed data mining systems: Issues and approaches, In *Advances in Distributed and Parallel Knowledge Discovery*, H. Kargupta and P. Chan (editors), AAAI/MIT Press, 2000.
[20] J. Ribeiro, K. Kaufman, and L. Kerschberg, Knowledge discovery from multiple databases. In: *Proceedings of KDD95.* 1995: 240-245.
[21] T. Shintani and M. Kitsuregawa, Parallel mining algorithms for generalized association patterns with classification hierarchy. In: Proc. of ACM SIGMOD, 1998: 25-36.
[22] K. Turinsky and R. Grossman, A framework for finding distributed data mining strategies that are intermediate between centeralized strategies and in-place strategies. In: *Proceedings of Workshop on Distributed and Parallel Knowledge Discovery at KDD-2000*, 2000: 1-7.
[23] G. Webb, Efficient search for association rules. In: *Proceedings of ACM SIGKDD*, 2000: 99-107.
[24] D.H. Wolpert, Stacked Generalization. *Neural Networks*, 5(1992): 241–259.
[25] S. Wrobel, An algorithm for multi-relational discovery of subgroups. In: J. Komorowski and J. Zytkow (eds.) *Principles of Data Mining and Knowledge Discovery*, 1997: 367-375.
[26] X. Wu and S. Zhang, Synthesizing High-Frequency Rules from Different Data Sources, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 2, March/April 2003: 353-367.
[27] X. Wu, C. Zhang and S. Zhang, Mining Both Positive and Negative Association Rules. In: *Proceedings of 19th International Conference on Machine Learning*, Sydney, Australia, July 2002: 658-665.
[28] J. Yao and H. Liu, Searching Multiple Databases for Interesting Complexes. In: *Proc. of PAKDD*, 1997: 198-210.
[29] S. Zhang, Knowledge discovery in multi-databases by analyzing local instances. PhD Thesis, *Deakin University*, 2001.
[30] C. Zhang and S. Zhang, *Association Rules Mining: Models and Algorithms*. Springer-Verlag Publishers in Lecture Notes on Computer Science, Volume 2307, p. 243, 2002.
[31] N. Zhong, Y. Yao, and S. Ohsuga, Peculiarity oriented multi-database mining. In: *Proceedings of PKDD*, 1999: 136-146.

Mercure: Towards an Automatic E-mail Follow-up System

Guy Lapalme[1] and Leila Kosseim[2]

***Abstract*— This paper discusses the design and the approach we have developed in order to deal effectively with customer e-mails sent to a corporation. We first present the current state of the art and then make the point that natural language tools are needed in order to deal effectively with the rather informal style encountered in the e-mails. In our project, called Mercure, we have explored three complementary approaches: classification, case-based reasoning and question-answering.**

***Index Terms*— Customer relationship management, automatic e-mail response, e-mail response management, text classification, case-based reasoning, question-answering**

I. Context of the problem

THE number of free-form electronic documents available and needing to be processed has reached a level that makes the automatic manipulation of natural language a necessity. Manual manipulation is both time-consuming and expensive, making Natural Language Processing (NLP) techniques very attractive. E-mail messages make up a large portion of the free-form documents available today and as e-mail becomes more and more popular, an automated e-mail answering service will become as necessary as an automated telephone service is today.

This paper discusses the use of natural language processing for dealing with e-mail automatically. Our work was developed in the context of e-mails regarding investors relations sent to a specific corporation but we believe that the approach can be applied to any Customer Relationship Management (CRM) application.

Although it is difficult to find reliable figures on the quality of online customer service (because of commercial interests and the fact that these figures are most often given by companies selling CRM systems) the following situation described in [1] seems to be typical:

> A recent Jupiter study[1] of the top 125 web sites found that 55% of customers expect accurate responses to e-mail within 6 hours, yet only 20% of companies are meeting their expectations. Forty-two percent of the sites never responded to the e-mails, took more than five days to respond to the questions, or had no e-mail address listed on their site.

[1] Jupiter Communications, "E-mail Customer Service: Taking control of Rising Customer Demand", 2000.

[1] RALI, DIRO, Université de Montréal, CP 6128, Succ. Centre Ville, Montréal (Québec) Canada, H3C 3J7 lapalme@iro.umontreal.ca

[2] CLaC Laboratory, Concordia University, 1455 de Maisonneuve Blvd. West, Montréal (Québec) Canada, H3G 1M8 kosseim@cs.concordia.ca

Given the fact that more than half of the people in the US and Canada now have an everyday access to e-mail, it is important for companies to make sure that their clients can use this medium for customer service inquiries. In the context of e-commerce, customers expect more access, continuous support and increased convenience and at the same time, they are less tolerant of poor response time, inaccurate answers or worse, non-responsiveness.

E-mail offers a number of advantages for customers compared to telephone calls: there are no tedious telephone menus and no waiting on the line for an available operator during business hours; with e-mail, the customer can formulate her request any time at her own pace and can continue her normal activities while waiting for the answer. The answer arrives in her usual mailbox and it can be kept for later reference. The customer no longer has to listen carefully to a verbal answer and take the risk of missing or forgetting critical information. However, because there is no immediate feedback between the operator and the customer, the later can never be certain that the request has been received. In addition, interaction between the operator and the customer is much more awkward and slow with e-mail than with a telephone call.

For an enterprise, using e-mail allows it to keep track of communications with its customers either for statistical or quality-control purposes. It is also possible to send more complete and complex instructions by e-mail and to include other media such as pictures, video or audio clips. In addition, it is cheaper to geographically or chronologically distribute e-mail answering to operators. On the other hand, e-mail is much less personal than direct contact with customers.

As described by Walker [25], e-mail should not be considered a substitute for all feedback from customers. In order to figure out *just when e-mail is really the right tool for the job* it is important to study this tool ogether with innovative ways to use it effectively.

II. Current approaches

The simplest level of e-mail answering systems is the so-called *auto-responder*[2]. These systems return a static document in response to an e-mail according to the presence of keywords in the subject or body of the message. As a variant, the user can fill a set of predefined fields in a web form to customize the response. An obvious drawback of these systems is that they do not analyze the content of free-form messages. The content of the text is reduced to a small set of keywords with no regards to the true meaning of the text.

[2] also known as *AR*, *infobots*, *mailbots* or *e-mail-on-demand*

More sophisticated types of e-mail responders are included in e-mail management systems, and can provide pre-written response templates for frequently asked questions. Slots are usually filled in with information extracted manually from the incoming mail, although some systems seem to perform the extraction automatically [19].

Some commercial systems such as Kana [17], RightNow [23] or XM-MailMinder [26] are aimed at optimizing the work flow of a call-center by keeping track of customer e-mails, helping representatives to answer by means of partially filled templates and providing productivity statistics on the answering process. However, to our knowledge, these systems do not use any NLP technology outside spell-checking and regular expression matching. Some systems also perform text classification (using learning techniques from annotated corpora or regular expressions) to categorize the incoming message into general pre-defined classes (e.g. requests, congratulations, complaints, ...). The e-mail can then be routed to the appropriate department or representative or, with specific categories, can even be answered automatically or deleted in the case of *spam*.

An early work on the automatic generation of appropriate answers to customer requests was performed by Coch [9], [10] who developed a system to generate answers to complaint letters from clients of La Redoute (a large French mail-order corporation). As letters were not in electronic format, the reading, the extraction and the decision was performed manually, but the production of a well-formed response was done automatically. Through a formal blind evaluation, Coch demonstrated that the best responses (according to specific criteria) are still the human-generated ones, but that the use of a hybrid template-based Natural Language Generation (NLG) system produced acceptable responses at a much faster rate.

III. Mercure

Bell Canada Enterprises (BCE) is a large Canadian corporation offering communication and entertainment services such as telephone, internet and television to private and commercial customers. To keep its competitive edge, its customer service must be efficient and cost-effective. In order to achieve this, BCE asked the Bell University Laboratories (BUL) to study the problem of e-mail follow-up in cooperation with the RALI (Recherche Applique en Linguistique Informatique[3]) laboratory. This has resulted in Mercure[4], a 4 year study, also funded by a Cooperative Research and Development grant from the National Science and Engineering Research Council (NSERC) of Canada.

After a preliminary study on a corpus of e-mails dealing with printer related problems [18], we focused on customer e-mails sent to a specific department at BCE: the investors relations department. This department receives and answers e-mails of current and potential investors sent to the address `investors.relations@bce.ca`. The e-mails are often requests for annual reports, press releases, but sometimes contain more complex financial questions such as values of stocks on specific dates, buying and selling plans, explanations about current events of the company; and also regard more routine issues such as address changes, lost of certificate, etc. Although the e-mail service is limited to administrative matters and that no judicial responsibility can be attributed to late or even false answers, timely and exact responses are essential for keeping good relations with investors.

[3]Applied Research in Computational Linguistics

[4]French name for Mercury, the roman god who was messenger of the other gods.

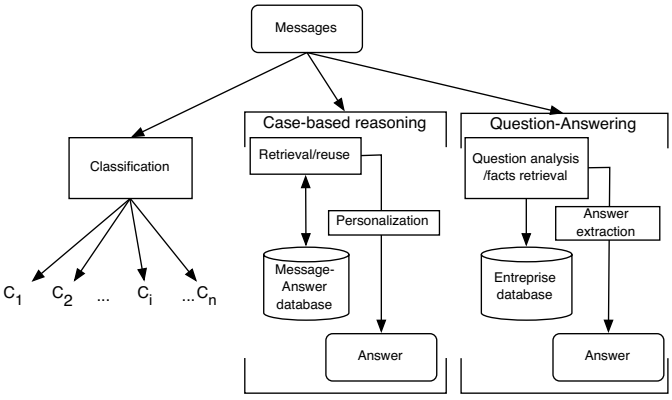


Fig. 6. Modules of the Mercure Project

In order to understand how e-mail is currently dealt with within BCE, we studied a preliminary corpus of more than 1000 e-mails sent to the investors relation department. The analysis showed that the e-mail varied considerably with regards to the level of difficulty required to analyze them: some e-mails were short and asked for a factual answer often found directly in a corporate documentation, while others were quite long and answering them required deeper research and information gathering from various sources. Because of this, we believed that a single technique could not suffice to deal with all e-mails, and we decided to try three complementary techniques in parallel and then to determine which one seems more appropriate given specific e-mail characteristics. Eventually, a combination of these techniques could be used in a real implementation. Figure III shows the three techniques explored in Mercure: text classification, case-based reasoning and question-answering. The following subsections will describe each technique in greater detail.

A. Classification

Classification of documents is a well known problem, but only recently has it been possible to use computers to separate texts into predefined categories according to their contents. The result of classification can be seen as a summary representation of the topic of a set of similar documents in order to ease the finding of related documents. Assigning a document to a certain class is not always a clear cut decision as a document may differ considerably from the others or could be assigned to more than one class. Text classification is typically performed using standard machine learning techniques and information retrieval term weighting schemes. Word distribution is a good feature for discriminating among categories and to classify a new document to its most appropriate category. Although much work has been done on the classification of

TABLE I
RESULTING CLASSIFICATION OF THE 818 SINGLE-PURPOSE MESSAGES OF OUR CORPUS.

Category	%	Description
dividend r.p.	5%	dividend reinvestment plan
stock split	5%	BCE stock split
dividend	5%	other questions about dividends
mailing list	7%	asking to be added or removed from a distribution list
report	17%	asking for annual or trimestrial reports
share price	29%	value of BCE stock
general	32%	other

newspaper articles through techniques such as *K nearest neighbors* [13], *naive Bayes* [15], decision trees such as CART [4] and ID3 [16]. Fewer projects have addressed the problem of e-mail classification [8], [11]. A notable exception is the classification of *spam*, which has attracted some interest in this problem and has even spurred an open-source project [14].

In the context of BCE, a seemingly simple problem is dealing with the intricacies of the contents of e-mail such as headers, citations, attachments, HTML parts, etc. that, in some cases, *hide* the text content and creates noise for the classifier. After removing this *noise*, Dubois [12] managed to extract the content of the e-mails in order to build a corpus of 1568 message and follow-up pairs sent between June 1999 and November 2000 to `investors.relations@bce.ca`. These e-mails were used by Dubois to study many types of classifiers (k nearest neighbors with k=10,20,30,40,50, naive Bayes network and Ripper) on different number of classes (5,10 and 22), with or without preprocessing (numeral and stop word removal or stemming, truncating words or not) and using different separation of corpus between training and validation sets. About 150 configurations have been tested with a success rate of about 50%. The main cause for errors was the noise brought by the fact that some messages dealt with more than one subject or were part of a multi-message exchange. So it was decided to work with only single-topic e-mails. With similar configurations as in the previous case and combination of them (210 in total), results raised to 90% for 5 categories, 80% for 10 categories and 67% for 22 categories. After studying the confusion matrices for all these cases, Dubois finally choosed the 6 categories (plus one *general*) shown in table III-A. With these categories, a success rate of about 80% was obtained on a 144 e- mail test set for March 2002, a period not contained in the learning set.

These results are adequate in the context of Mercure because e-mail of some of these classes (*dividend r.p.* and *mailing list*) are already being forwarded to people outside of BCE. Messages of the *report* category are answered by simply mailing the desired report.

B. Case-Based Reasoning

The second approach we are investigating is the application of textual case-based reasoning (CBR) techniques to generate responses to incoming email messages. This CBR module exploits a corpus of email messages comprising requests from investors and their corresponding responses from financial analysts. Case-based reasoning is similar in spirit to the way humans reuse (and adapt) previous e-mails for answering new requests. The design of a CBR email response system relies on a corpus of previously answered messages, a resource that is representative of the domain of discourse and of the various problems tackled during email exchanges. The *search and adapt* reasoning scheme then offers a natural mapping to the two phases of email response, i.e. the analysis of incoming requests and the synthesis of relevant responses. Presented from a client perspective, the CBR module attempts to reuse messages in the SENT mailbox of the analyst's email software to suggest responses to new messages incoming in the INBOX. Our processing is divided into three main phases (retrieval of cases, reuse of cases and personalization of the answer). Each step is now described below and has been implemented in a prototype Java-based mail client.

1) Retrieval of cases: This phase compares a new message with the ones previously received, in order to find a similar one and reuse its answer. During our initial experimentation, the similarity between messages was established based on the comparison of a tf.idf (term frequency $\times$ inverse document frequency) vectorial representation of the message content. Using a cosine function to compute global similarity provides a precision of approximately 57.9%. This is similar to the results of comparable experiments with FAQs [7]. However, the nature of our cases can be exploited to improve some aspects of the retrieval phase. As the selection of wrong answers requires additional manipulation by the user of the system, it is important to optimize the ranking of the most relevant(s) case(s) to ensure the production of a relevant response.

For improving the performance of the retrieval phase, we first considered the classical word relationships but it required an exact correspondence of words (or key-phrases or ngrams). To overcome this constraint, some authors [6], [7] have made use of existing linguistic resources (e.g. thesaurus) to establish the semantic similarity of different words that have related meanings. This approach does not transpose well to our problem as, to our knowledge, no domain specific resources are available.

Since textual responses provided by a limited number of analysts are more similar (based on word distribution) than requests sent by many different investors, we conjectured that similarity should be more easily established when the textual responses are also taken into account during the retrieval phase. We combined both of the above possibilities into a single scheme. A textual case can be seen as the linguistic *conversion* of a textual problem into a corresponding textual solution. The case base then corresponds to a mapping from a *request* language (problem) to a *response* language (solution). The finding of associations, captured as co-occurrences, provides indications that the occurrence of problem words increases the likelihood of the presence of some other words in the solution. To obtain the co-occurences, we collect the count of all pairs of words coming respectively from the requests and their corresponding responses, and we select the most significant ones based on the mutual information metric [21].

The approach we are currently using of inserting the associ-

ations in the retrieval phase is inspired from query expansion techniques. The incoming problem description (the investor's request) is expanded into a vector of response terms provided by the lists of co-occurrences. Similarity of the cases then corresponds to the weighted sum of both problem and solution vector cosine. Experimentation [20] conducted on 102 test requests indicates that the expansion scheme slightly improves the overall precision (62.0% vs. 57.9%) of the retrieval phase and preserves the rank of the first pertinent solution in the similarity list (2.01 vs. 1.96). The most significant improvement has been observed for the test messages where the response is not directly addressing the request (e.g. redirection to a generic web site address following the request of specific documents or financial information). For this category of message, the precision is almost doubled (80.1% vs. 51.0%) and the average rank is reduced to a very good level (1.33 vs. 2.38). For the other messages, the precision is mostly preserved but we observed some degradation for the routine messages as the expansion scheme introduces some noise in the internal representation of the textual cases. This result is however interesting as responses are built from a limited number of the most highly ranked cases (usually the first one). And, most importantly, we expect that the selection of a judicious trade-off between request and solution similarities will bring further improvement.

2) Reuse of previous cases: Our application presents strong incentives to implement some adaptations of previous responses. While complete reformulation of past textual responses for diverse situations is beyond the capability of current CBR and NLP techniques, some of these techniques can nevertheless help to personalize past messages and preserve the relevance of cases with the context of the new incoming request. In the CBR literature, case adaptation (i.e. case reuse) has exclusively been conducted for structural cases and mostly corresponds to modifying the values of pre-selected solution features. In a textual setting like our email response domain, such a scheme is rather difficult to implement, as the textual solutions are not structured. Therefore, prior to the modification of the content of the messages, we need to determine what portions of the responses are good candidates for modification. Given a new message and some past solutions selected during the retrieval phase, we have implemented the reuse of textual cases as a three-step process:

1) identification of passages for determining the text portions that are applicable in the context of the new incoming request. Statistical distributions, captured as word alignments [5], can be used for this task;
2) message personalization that determines what text portions are to be modified;
3) pruning and substitution for removal of irrelevant passages and the substitutions of the portions to be personalized. In NLP, this corresponds to a query-relevant summarization process [3], more specifically to the condensation of a text based on the terms of a request.

3) Personalization of the messages: Personalization of messages refers to the capacity to detect some factual information in the messages and to substitute them in the responses. This includes, for instance, names of companies, individuals, financial factors, dates and time references. These expressions correspond to named entities and can be identified using information extraction techniques (IE). IE techniques identify, using either rule patterns or statistical models, information from textual documents to be converted into a template-based representation. As we did during the first phase of the project, we make use of extraction patterns and lexicons (lists of company names, titles, acronyms and frequent financial terms).

Substitutions of these entities are partly conducted using a rule-based approach. Replacement of individual names and companies is based on the roles of the messages entities. The role is determined by the type of patterns used during extraction, mostly based on the part-of-speech and the terms preceding/following the entities. For instance, expressions like *"Sincerely, John Smith"*, *"to purchase Nortel shares"*, *"registered with Montreal Trust"*, could provide indications of the message sender, subsidiary company and financial institution respectively. However, as the Investor Relations domain does not offer much predictability, the elicitation of domain rules for numeric information (dates, price, factors?) remains difficult and such substitutions rely mostly on the user.

C. Question-Answering

Many of the e-mails sent to corporations are asking for information and can be considered as questions from customers to which representatives should answer in the best possible way. The third technique used is based on Question-Answering (QA) technology: the task of finding an exact answer to a natural language question [24] in a large set of documents. The question type is determined by the presence of trigger phrases (e.g. *where*, *how many*, *how much*), which indicate the type of the answer required (e.g. *location, number, money*). Information retrieval is typically performed to identify a subset of the documents and a set of passages that may contain the answer. Named entities are then extracted from these passages and semantically tagged and the string containing the best scoring entity is retained as the answer. Within Mercure, we have developed Quantum [22], a *traditional* QA system with which we participated in the QA-track of TREC and that will be used as a basis for our work in e-mail answering.

QA differs from e-mail answering in several aspects. Generally speaking, e-mail answering involves *analyzing* a longer text and *formulating* a linguistically-motivated answer, while QA takes a short and explicit question as input and focuses on *locating* the answer. Issues in discourse analysis and generation must therefore be addressed in e-mail answering, but not in QA. In addition, questions, at least in systems participating to the TREC evaluations, are restricted to specific types such as *who, why, where,* ... but pertain to an unrestricted discourse domain. On the other hand, in e-mail answering, the questions are of unrestricted type, but the discourse domain is typically restricted. E-mail answering thus involves finding passages from the textual knowledge base that best relate to the incoming message and sending the passages as is to the user. This is the avenue currently being pursued by Luc Blanger[2] in his Ph.D. thesis.

IV. Transfer to the industry

In order to make sure that the technology we developed in our lab could be transferred to the operational context of BCE, we installed a mirror mail server with the same hardware and software configuration as the one used by BCE. We also made arrangements to receive a copy of all e-mails sent to investors relations at BCE and this enabled us to build a dynamic corpus of e-mails which was used for testing: these new e-mails deal with the same domains as the ones used for developing the system. A version of the classifier has been installed in the BCE mail server but administrative delays and change of personnel did not allow a complete integration into the answering process. The CBR and Question-Answering modules are being developed separately and will eventually be integrated into the mail server.

V. Conclusion and future work

In this paper we have described the research conducted within the Mercure project, aimed at the automatic follow-up of e-mail messages. The work was performed specifically with a corpus of e-mails from the investors relations department of Bell Canada Enterprises. As the e-mails were not homogeneous in their textual characteristics, we explored three complementary approaches: text classification, case-based reasoning and question-answering. Our experience with e-mail classification was not very fruitful. As the classes considered were very much related, the standard word distribution approach showed insufficient discrimination power. However, it would be interesting to compare our results with human classification to have an upper bound measure of what we can hope to achieve. This would allow us to evaluate whether the approach needs to be modified or if the task is simply too difficult. The 2 other approaches are still under development. The case-based reasoning module seems promising and the research performed so far seems to show that an important number of messages can be answered using this technique. Finally, the question-answering approach still needs more work, especially to identify the question in the texts.

Once the case-based reasoning and the question-answering modules are in place, we plan to evaluate each approach on different sets of e-mails so as to measure how appropriate each approach is as a function of specific e-mail characteristics such as e-mail length, category, etc. This will allow us to combine the three approaches either by running then in parallel and combining their result, or by using one approach and revert to another if the previous one is unable to produce an appropriate answer with enough confidence.

Acknowledgments

The work described in this article is a joint work with the following graduate students: Stéphane Beauregard, Luc Plamondon, Luc Lamontagne, Luc Bélanger and Julien Dubois. We are grateful to our colleagues from RALI for their relevant comments. This work was supported by a grant from Bell University Laboratories and a Cooperative Research and Development grant from NSERC.

References

[1] Banter Inc. Natural language engines for advanced customer interaction, 2001. `http://www.realmarket.com/required062801.html`.

[2] Luc Bélanger. Le traitement automatisé des courriels pour les services aux investisseurs: une approche par la question-réponse. Technical report, Département d'informatique et RO - Université de Montréal, 2003.

[3] A. Berger and V. Mittal. Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 294–301, Hong-Kong, 2000.

[4] L. Breiman, J.H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth Int. Group, 1984.

[5] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, R. Jelinek, F.and Mercer, and P. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[6] S. Brüninghaus and K. Ashley. Bootstrapping case base development with annotated case summaries. In *Proceedings of ICCBR-99, Lecture Notes in Computer Science 1650*, pages 59–73. Springer Verlag, 1999.

[7] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. *AI Magazine*, 18(2):57–66, 1997.

[8] S. Busemann, S. Schmeier, and R. Arens. Message Classification in the Call Center. In *Proceedings of ANLP-2000*, pages 159–165, Seattle, 2000.

[9] J. Coch. Evaluating and comparing three text-production techniques. In *Proceedings of COLING-96*, Copenhagen, Dannemark, 1996.

[10] J. Coch and J. Magnoler. Quality tests for a mail generation system. In *Proceedings of Linguistic Engineering*, Montpellier, France, 1995.

[11] W. Cohen. Learning rules that classify e-mail. In *Proceeding of the 1996 AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[12] Julien Dubois. Classification automatique de courrier électronique. Master's thesis, Université de Montréal, 2002.

[13] R. O. Duda and P.E. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, 1973.

[14] John Graham-Cumming. Popfile - automatic email classification, 2003.

[15] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorisation. In D.H. Fisher, editor, *Proceedings of the 14th International Conference on Machine Learning*, pages 143–151, San Francisco, 1997. Morgan Kaufmann.

[16] Quinlan J.R. Induction on decision trees. *Machine Learning*, 1(1):81–106, 1986.

[17] www.kana.com. 1999.

[18] Leila Kosseim, Stéphane Beauregard, and Guy Lapalme. Using information extraction and natural language generation to answer e-mail. *Data & Knowledge Engineering*, 38:85–100, 2001.

[19] Y. Lallement and M. Fox. Interact: A Staged Approach to Customer Service Automation. In H. Hamilton and Q. Yang, editors, *Canadian AI 2000*, LNAI 1822, pages 164–175, Berlin, 2000. Springler-Verlag.

[20] Luc Lamontagne, Philippe Langlais, and Guy Lapalme. Using statistical word associations for the retrieval of stronglly-textual cases. In *Florida Artificial Intelligence Research Science (FLAIRS) 2003*, page 7 pages, St-Augustine, Florida, 2003.

[21] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[22] Luc Plamondon, Leila Kosseim, and Guy Lapalme. The QUANTUM question answering system at trec-11. In E.M. Vorhees and D.K. Harman, editors, *The Eleventh Text REtrieval Conference (TREC-2002)*, pages 670–677, Gaithersburg, MD, November 2002. NIST.

[23] RightNow Technologies. Revelation Knowledge Engine, 2002. `http://www.rightnow.com`.

[24] *Proceedings of the Tenth Text REtrieval Conference (TREC-X)*, Gaithersburg, Maryland, 2001.

[25] David Walker. Automation woes widen the email expectations gap. `http://www.shorewalker.com/pages/email_expectations-1.html`.

[26] XtraMind Technologies GmbH. XM-MailMinder, 2002. `http://www.xtramind.com`.

Cross-Language Information Retrieval

Jian-Yun Nie[1]

***Abstract*—A research group in University of Montreal has worked on the problem of cross-language information retrieval (CLIR) for several years. A method that exploits parallel texts for query translation is proposed. This method is shown to allow for retrieval effectiveness comparable to the state-of-the-art effectiveness. A major problem of this approach is the unavailability of large parallel corpora. To solve this problem, a mining system is constructed to automatically gather parallel Web pages. The mining results are used to train statistical translation models.**

When a query is translated word by word, the accuracy may be low. In order to increase the translation accuracy, compound terms are extracted and incorporated into the translation models, so that compounds can be translated as a unit, rather than as separate words. Our experiments show that this can further increase the CLIR effectiveness.

I. Introduction

INFORMATION retrieval (IR) tries to identify relevant documents for an information need, expressed as a query. The problems that an IR system should deal with include document indexing (which tries to extract important indexes from a document and weigh them), query analysis (similar to document indexing), and query evaluation (i.e. matching the query with the documents). Each of these problems has been the subject of many studies in IR.

Traditional IR identifies relevant documents in the same language as the query. This problem is referred to as monolingual IR. Cross-language information retrieval (CLIR) tries to identify relevant documents in a language different from that of the query. This problem is more and more acute for IR on the Web due to the fact that the Web is a truly multilingual environment. In addition to the problems of monolingual IR, CLIR is faced with the problem of language differences between queries and documents. The key problem is query translation (or document translation). This translation raises two particular problems [6]: the selection of the appropriate translation terms/words, and the proper weighting of them. In the last few years, researchers have worked on these problems intensively. Three main techniques for query translation have been proposed and tested:

- With an on-the-shelf machine translation (MT) system;
- With a bilingual dictionary;
- Or with a set of parallel texts.

[1] Département d'Informatiqueet Recherche opérationnelle, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, H3C 3J7 Canada, nie@iro.umontreal.ca

The first two approaches are quite straightforward. We will not give details about them. Our research efforts have been concentrated on the third approach. This approach is promising because it does not require extensive manual preparation (in comparison with the construction of an MT system); and its translation is usually more appropriate than with a bilingual dictionary.

The major advantages of this approach are the following ones:

The training of a translation model can be completely automatic. No (or little) manual preparation is required.

The resulting translation model reflects well the word usage in the training corpus. This offers the possibility to train specialized and up-to-date translation models.

In this paper, we will describe our approach to CLIR based on parallel texts, as well as some experiments. The paper will be organized as follows. In Section II, we will first describe briefly the training process of statistical translation models on a set of parallel texts. Then we will describe in Section III the IR system we use for our experiments. Section IV describes our experiments with the translation models trained on a manually prepared parallel corpus. Section V describes our approach to mining parallel Web pages, as well as their utilization for CLIR. Section VI presents our utilization of compound terms in CLIR. Finally, we present our conclusions in Section VII.

II. Training Statistical Translation Models on Parallel Texts

Let us first describe briefly the training of statistical translation models on a set of parallel texts. These models will be used in our experiments.

Statistical translation models are trained on parallel texts. A pair of parallel texts is two texts which are translation one of the other. Model training tries to extract the translation relationships between elements of the two languages (usually words) by observing their occurrences in parallel texts. Most work on the training statistical translation models follows the models (called IBM models) proposed by Brown *et al*. [1]. In our case we use the IBM model 1. This model does not consider word order in sentences. Each sentence is considered as a bag of words. Any word in a corresponding target sentence is considered as a potential translation word of any source word. This consideration is oversimplified for the purpose of machine translation. However, for IR, as the goal of query translation is to identify the most probable words without considering the syntactic features, this simple

translation model may suffice.

In order to train a translation model, parallel texts are usually decomposed into aligned sentences, i.e. for each sentence in a text, we determine its translation sentence(s) in the other language. The primary goal of producing sentence alignment is to reduce the scope of translation relationships between words: instead of considering a word in a source text to correspond potentially to every word in the target text, one can limit this relationship within the corresponding sentences. This allows us to take full advantage of the parallel texts and to produce a more accurate translation model.

A. Sentence Alignment

Sentence alignment tries to create translation relationships between sentences. Sentences are not always aligned into 1:1 pairs. In some cases, one sentence can be translated into several sentences, and the sentence may even be deleted or a new sentence may be added in the translation. This adds some difficulties in sentence alignment.

Gale & Church [5] propose an algorithm based on sentence length. It has been shown that this algorithm can successfully align the Canadian Hansard corpus (the debates in the Canadian House of Commons in both English and French), which is rather clean and easy to align. However, as pointed out by Simard *et al.* [12] and Chen [3], while aligning more noisy corpora, the methods based solely on sentence length are not robust enough to cope with the above-mentioned difficulties. Simard *et al.* proposed a method that uses lexical information, cognates, to help with alignment [12].

Cognates are pairs of tokens of different languages, which share obvious phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. Examples are generation/génération and financed/financé for English/French. In a wider sense, cognates can also include numerical expressions and punctuation. Instead of defining a specific list of cognates for each language pair, Simard *et al.* gave language-independent definitions on cognates. Cognates are recognized on the fly according to a series of rules. For example, words starting with 4 identical letters in English and French are considered as cognates.

Another method incorporates a dictionary [3]. The translations contained in the dictionary serve as cues to sentence alignment: a sentence is likely to align with another sentence if the latter contains several dictionary translations of the words of the former.

In our implementation, we use the approach of Simard *et al.* [12].

B. Model Training

The principle of model training is: in a set of aligned sentences, if a target word f often co-occur with a source word e in the aligned sentences, then there is a high chance that f is a translation of e, i.e. the translation probability $t(f|e)$ is high. The training algorithm uses dynamic programming to determine a probability function $t(f|e)$ such that it maximizes the expectation of the given sentence alignments (see [1] for details).

We briefly describe the training for IBM model 1 as follows.

The translation probability function t is determined such as to maximize the probability of the given sentence alignments A of the training corpus. Suppose a sentence alignment $\mathbf{e}\leftrightarrow\mathbf{f}$, and that the sentences $\mathbf{e}$ and $\mathbf{f}$ are composed of set of words as follows:

$$\mathbf{e}=\{e_1, e_2, e_3, ..., e_l\},$$
$$\mathbf{f}=\{f_1, f_2, f_3, ..., f_m\}$$

where l and m are respectively the length of these sentences. Then the function t is determined as follows:

$$\begin{aligned} t &= \arg\max_t p(A) \\ &= \arg\max_t \prod_{\mathbf{e}\leftrightarrow\mathbf{f}} p(\mathbf{f}\mid\mathbf{e}) \\ &= \arg\max_t \prod_{\mathbf{e}\leftrightarrow\mathbf{f}} \varepsilon(1+l)^{-m} \prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j \mid e_i) \end{aligned}$$

where ε is the probability that an English sentence of length l can be translated into a French sentence of length m, and $t(f_j|e_i)$ the word translation probability of e_i by f_j.

The probability t can be determined by applying the iterative EM (Expectation maximization) algorithm. We do not give details here. Interested readers can refer to [1].

IBM model 1 considers every word in the target sentence to be equivalently possible translation of any word in the source sentence, regardless to their position and to the "fertility" of each word (e.g. an English word may be translated by one or more French words). It is obvious that the translation model does not learn syntactic information from the training source and thus cannot be used to obtain syntactically correct translations. However, the model is able to determine the word translation probability t between words, and this fits the need of cross-language information retrieval of finding out the most important translation words.

III. IR SYSTEM

In our experiments we use the SMART system. SMART is an IR system, developed at Cornell University [2]. The indexing process considers every token as an index. Indexes are weighted according to the *tf*idf* weighting scheme[2]. This is a common way to weigh the importance and specificity of a term in a document. The principle is as follows: 1) the more a word occurs in a document, the more it is important. This is the *tf* factor. On the other hand, the more there are documents containing the word, the less the word is specific to one particular document. In other words, the word does not allow distinguishing a document from the others. Therefore, the weight of the word is lowered. This is the *idf* factor. More precisely, the two factors are measured as follows:

[2] tf = term frequency, and idf = inversed document frequency.

$$tf(t, D) = \log(freq(t, D) + 1);$$
$$idf(t) = \log(\frac{N}{n(t)})$$

where *freq*(*t, D*) is the frequency of occurrences of the word/term *t* in the document *D*; *N* is the total number of documents in the collection; *n*(*t*) is the number of documents containing *t*.

The retrieval process follows the vector space model [2]. In this model, a vector space is defined by all the tokens (words or terms) encountered in the documents. Each word/term represents a distinct dimension in this space. Then a document, as well as a query, is represented as a vector in this space. The weight in a dimension represents the importance of the corresponding word/term in the document or query (the *tf***idf* weight). The degree of correspondence between a document and a query is estimated by the similarity of their vectors. One of the commonly used similarity measures is as follows:

$$sim(D, Q) = \frac{\sum_i d_i \times q_i}{\sqrt{\sum_i d_i^2 \times \sum_i q_i^2}}$$

where d_i and q_i are respectively the weights of a term in the document *D* and in the query *Q*.

IV. Experiments with the Hansard models

There are a few manually constructed parallel corpora. The best known is the Canadian Hansard, which contains the debates of the Canadian parliaments during 7 years, in both French and English. It contains dozens of millions words in each language. Such a parallel corpus is a valuable resource that contains word/term translations. Our first experiments are carried out with translation models trained on the Hansard corpus- we call the resulting models the Hansard models.

We used two test collections developed in TREC[3], one in English (AP) and the other in French (SDA). Both collections contain newspaper articles. The SDA contains 141,656 documents, and AP 242,918 documents. We use two sets of about 30 queries, available in both French and English. These queries have been used in TREC6 and TREC7 for French-English CLIR. The queries have been manually evaluated (i.e. we know their relevant documents). Table I shows the CLIR effectiveness obtained with these translation models. F-E means using French queries to retrieve English documents, i.e. the French queries are first translated into English, then the English translation is used to match the documents. In all our experiments, we select the 25 most probable translation words as the "translation" of a query.

In Table I, the effectiveness is measured by average precision, i.e. the average of the precisions over 11 points of recall. This is a standard measure used in IR. We also show the percentage of the CLIR effectiveness with respect to the monolingual IR effectiveness (%mono). In comparison with the state-of-the-art effectiveness, which is usually around 80-90% of the monolingual effectiveness (see the reports of TREC at http://trec.nist.gov), the results we obtained are quite comparable.

TABLE I.
AVERAGE PRECISION USING HANSARD MODEL

	F-E (%mono)	E-F (%mono)
Trec6	0.2166 (74.8%)	0.2501 (67.9%)
Trec7	0.3124 (97.6%)	0.2587 (93.6%)

V. Mining of parallel Web pages

A major problem to use parallel texts is often the unavailability of large parallel corpora. In order to obtain such corpora, we constructed a mining system – PTMiner [4] – to automatically gather parallel Web pages.

Although many parallel Web pages exist on the Web, it is not obvious to identify them and to confirm that a pair of pages is truly parallel. In our mining approach, we exploit several heuristic features. For example, if an English page points to another page with an anchor text "French version" or "version française", this is a useful indication that the second page is a French version of the first page. Although these indications are not fully accurate, and they can produce errors, we will show later in our experiments that a noisy parallel corpus is still useful for query translation in CLIR.

In the following subsections, we will briefly describe our mining approach.

A. Automatic Mining

Parallel web pages often are not published in isolation. Most of the time, they are connected in some way. For example, Resnik [11] observed that parallel Web pages often are referenced in the same parent index web page. In addition, the anchor text of such links usually identifies the language. For example, if a home page "`index.html`" contains links to both English and French versions of the next page, and that the anchor texts of the links are respectively "`English version`" and "`French version`", then the referenced pages are parallel. In addition, Resnik assumes that parallel Web pages have been indexed by large search engines existing on the Web. Therefore, in his approach, a query of the following form is sent to Alta Vista in order to first retrieve the common index page:

```
anchor: english  AND  anchor: French
```

Then the referenced pages in both languages are retrieved and considered to be parallel pages.

We notice that only a small number of web sites are organized in this way. Many other parallel pages do not satisfy this condition. Our mining strategy uses different criteria. In addition, we also incorporate an exploration process (host crawler) in order to discover more web pages

[3] TREC: Text Retrieval Conference, a series of conferences aiming to test IR systems with large document collections. See http://trec.nist.gov/

that have not been indexed by the existing search engines.

Our mining process is separated into two main steps: first identify as many candidate parallel pages as possible, then verify external features and contents to determine if they are parallel. Our mining system is called PTMiner (for Parallel Text Miner). The whole process is organized into the following steps:

1. Determining candidate sites – This step tries to identify the Web sites where there may be parallel pages.
2. File name fetching – It identifies a set of Web pages from each Web page that are indexed by search engines.
3. Host crawling – It uses the URLs collected in the last step as seeds to further crawl each candidate site for more URLs.
4. Pair scanning by names – It pairs the Web pages according to the similarity of their URLs.

IDENTIFICATION OF CANDIDATE WEB SITES

To determine candidate sites, we assume that a candidate site contains at least one page that refers to another version of the page, and the anchor text of the reference clearly identifies the language. For example, an English Web page contains a link to the French version, and the anchor text is "French version", "in French", "en français" and so on. So to determine the candidate sites, we send a particular request to search engines asking for English pages that contain a link with an anchor text identifying another language such as:

```
anchor: french version, [in french, ...]
language: English
```

The host addresses we extract from the resulting Web pages correspond to the candidate sites.

FILE NAME FETCHING

To search for parallel pairs from each candidate site, PTMiner first asks the search engines for all the Web pages from this site they have indexed. This is done by a query of the following form:

```
host: <hostname>
```

However, a search engine may not index all the Web pages on a site. To obtain a more complete list of URLs from a site, we need to explore the sites more thoroughly by a host crawler.

HOST CRAWLING

A host crawler is slightly different from a Web crawler or a robot [10] in that a host crawler only exploits one Web site. A breadth-first crawling algorithm is used in this step. The principle is that if a retrieved Web page contains a link to an unexplored document on the same site, this document is added to a list that will be explored later. This crawling step allows us to obtain more web pages from the candidate sites.

PAIR SCANNING BY NAMES

We observe that many parallel pages have very similar file names. For example, an English web page with the file name "index.html" often corresponds to a French translation with the file name "index_f.html", "index_fr.html", and so on. The only difference between the two file names is a segment that identifies the language of the file. This same observation also applies to URL paths. In some cases, the two versions of the web page are stored in two different directories, for example, `www.asite.ca/en/afile.html` vs. `www.asite.ca/fr/afile.html`. So in general, a similarity in the URLs of two files is a good indication of their parallelism. This similarity is used to make a preliminary selection of candidate pairs.

FILTERING AFTER DOWNLOADING

The remaining file pairs are downloaded for further content verification according to the following criteria:.

- Length of the pages: A pair of parallel pages usually has similar file lengths. A simple verification is then to compare the lengths of the two files. Note that the length ratio changes between different language pairs.
- HTML structure: Parallel web pages are usually designed to look similarly. This often means that the two parallel pages have similar HTML structures. Therefore, the similarity in HTML tags is another filtering criterion.
- The pair-scanning criterion we used only exploits the name similarity of parallel pages. This is not a fully reliable criterion. Files with a segment "en_" may be not in English. Therefore, a further verification is needed to confirm that the files are in the required languages. In our system, we use the SILC4 system for an automatic language and encoding identification.

With PTMiner, we have been able to collect several parallel corpora from the Web. Table II shows some of them.

TABLE II.
SIZES OF THE WEB CORPORA

	FR-EN		DE-EN		IT-EN	
# Text Pairs	18 807		10 200		8 504	
Raw data (MB)	198	174	100	68	50	77
Cleaned data (MB)	155	145	66	50	35	50

In our further description, we will concentrate on the French-English pair.

B. CLIR With the Web Models

Translation models are trained on the set of parallel Web pages as described in Section 2, except that some preprocessing has to be performed on these pages in order to remove HTML tags. Once translation models (in both directions) are trained, they are used to produce 25 most probable translation words that are considered as the translation of a query. Table III describes the CLIR

[4] See http://www-rali.iro.umontreal.ca/ProjetSILC.en.html

effectiveness with the Web models.

TABLE III.
AVERAGE PRECISION USING WEB MODEL

	F-E (%mono)	E-F (%mono)
Trec6	0.2103 (72.6%)	0.2595 (70.4%)
Trec7	0.2380 (74.3%)	0.1975 (71.5%)

In comparison with the Hansard model, we see that the Web models perform slightly worse. However, considering the noise that this training corpus may contain, this effectiveness is quite good. It is still close to the state-of-the-art effectiveness. This test shows that the automatically mined parallel Web pages are greatly useful for CLIR.

VI. INCORPORATING COMPOUND TERMS IN TRANSLATION MODELS

In the previous approach, parallel texts have been exploited to find translations between single words. The most obvious problem we can see is that by taking words one by one, many of them become ambiguous. The translation model will then suggest several translations corresponding to different meanings of the word. For example, the word "information" (in French) will have many possible translations because 1) the word denotes several meanings; 2) it appears very frequently in the parallel corpus. Among the possible translations, there are "information", "intelligence", "espionage", etc. However, if the term we intend to translate is "système d'information" (information system), and if the term is translated as a whole, then many of the meanings of "information" can be eliminated. The most probable translation of this term will be the correct term "information system". Through this example, we can see that a translation model that integrates the translation of compound terms can be much more precise. This is the goal of our utilization of compounds during query translation.

To do this, we have to train a translation model that incorporates compound terms as additional translation units to words. So compound terms are first extracted from the training parallel corpus, and added to the original sentences. Then the same translation process is launched. The resulting model contains the translations for both single words and compound terms.

To identify compound terms, we use both a large terminology database containing almost 1 million words and terms, and an automatic extractor of compound terms. The extractor uses syntactic structures, together with a statistical analysis. First, word sequences corresponding to predefined syntactic templates are extracted as candidates. If the frequency of occurrences of a candidate is above a certain threshold, then the sequence is considered as a compound term.

The first problem is the definition of the syntactic templates. This is done manually according to the general knowledge on syntactic structures of a language. Usually the extraction is restricted to noun phrases. For example, the following template is used in the tool we used - Exterm:

```
((NC|AJ) )*((NC|AJ)|NC PP) ((NC|AJ) )*NC
```

where NC means a common noun, AJ an adjective, and PP a preposition.

Of course, a POS (Part-Of-Speech) tagging is necessary in order to recognize the syntactic category of each word. The tagger we used is a statistical tagger trained on the Penn Treebank[5]. It tries to determine the most probable syntactic categories that fit the best the words of a sentence. Details on the training of such a tagger can be found in [7].

All the terms and words in documents, queries and the training parallel corpus are submitted to a standardization process on words, as follows:

- Nouns in plural are transformed into singular form (e.g. systems → system);
- Verbs are changed into infinitive form (e.g. retrieves → retrieve, retrieving → retrieve);
- Articles in a term is removed (e.g. the database system)

For example, the expression "adjusted the earnings" will be transformed into "adjust earning".

Once a compound term is recognized in a document or a query, it is added into the document or query. For example, consider a preprocessed text as follows:

```
arm dealer prepare relief supply to soviet union
```

From this segment, we can extract two stored terms "arm dealer" and "soviet union" So the following terms are appended to the original text:

```
arm_dealer soviet_union
```

Once compound terms are extracted from the training texts, the corpus is submitted to the training process of translation models described in Section 2. However, as compounds are considered as units of the texts, the resulting translation models will also contain translations for the compounds, which are usually more accurate than their word-by-word translations.

A. Experiments on CLIR

Table IV shows the CLIR results with both types of translation model. These results are obtained on the same document collection as the one used earlier, but the query set is different.

In these experiments, we separate single words and compound terms into two separate vectors. SMART has the flexibility of building multiple vectors for a document and for a query. Then the global similarity between the document and the query is determined by the weighted sum of the similarities between the vectors. One can assign a relative weight to different vectors of the query to balance their importance in the global similarity.

In our experiments, we tested several values for the relative weights of the single-word vector and the compound-term vector. The above results are obtained with the relative

[5] http://www.cis.upenn.edu/~treebank/home.html

importance of 0.3 to the compound-term vector, and 1 to the single-word vector. This assignment gives the best result.

We can see a great improvement in CLIR effectiveness once the translation model incorporates compound terms, especially for the F-E case. We have not applied the same approach to the Web corpus. However, we could expect similar improvements with the Web corpus when compound terms are incorporated.

TABLE IV.
THE CLIR EFFECTIVENESS WITH DIFFERENT MODELS.

	Word	Compounds (change)
F-E on AP data set	0.1465	0.2591 (+76.86%)
E-F on SDA data set	0.2257	0.2860 (+26.72%)

VII. CONCLUSIONS

In this paper, we described an approach based on parallel texts that has bee used for CLIR at University of Montreal. Globally, our experiments show that the statistical translation models trained on parallel texts are highly useful for CLIR. They can achieve comparable effectiveness to the state-of-the-art approaches. Our further tests with the parallel Web pages mined automatically show that we can arrive at a reasonable level of effectiveness despite the relatively high rate of noise in the training parallel Web pages. This series of experiments show that our method based on parallel Web pages is suitable for CLIR.

Nevertheless, we also observe several aspects that require improvements:

- We encounter problems for translating proper names. Proper names are often treated as unknown words, and are added into the translation as it is. For some names, the spellings in all the European languages are the same, which does not raise particular problems. For some others with different spellings (e.g. “Bérégovoy” in French, but “Beregovoy” in some English documents), this simple approach does not solve the problem.
- Translation by common but non stop- words: Very often, among the top translation words, the common words such as “prendre” and “donner” (“take” and “give” in French) appear with quite strong probability.
- The mined parallel Web pages contain a certain amount of noise. To improve the translation accuracy, a further filtering of noise is necessary.
- We are currently investigating on these problems.

REFERENCES

[1] P. F. Brown, S. A. D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. Computational Linguistics, vol. 19, pp. 263-312 (1993).

[2] Buckley, C. (1985) Implementation of the SMART information retrieval system. Cornell University, Tech. report 85-686.

[3] Chen, S. F. Aligning sentences in bilingual corpora using lexical information. Proc. ACL, pp. 9-16, 1993.

[4] J. Chen, J.Y. Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proc. ANLP, pp. 21-28, Seattle (2000).

[5] W. A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19: 1, 75-102 (1993).

[6] G. Grefenstette. The Problem of Cross-Language Information Retrieval. In *Cross-language Information Retrieval*. Kluwer Academic Publishers. pages 1-9, 1998

[7] C. Manning, H. Shultze, *Fundamentals of Statistical Natural Language Processing*, MIT Press, 1999

[8] J.Y. Nie, P. Isabelle, M. Simard, R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, *ACM-SIGIR conference*, Berkeley, CA, pp. 74-81(1999).

[9] J.Y. Nie, J.F. Dufort, Combining Words and Compound Terms for Monolingual and Cross-Language Information Retrieval, *Information 2002*, Beijing, July 2002.

[10] Prosise J., Crawling the Web, A guide to robots, spiders, and other shadowy denizens of the Web, PC Magazine - July 1996 (http://www.zdnet.com/pcmag/issues/1513/pcmg0045.htm).

[11] Resnik, Philip (1998) Parallel stands: A preliminary investigation into mining the Web for bilingual text, *AMTA'98*, *Lecture Notes in Artificial Intelligence, 1529,* October.

[12] M. Simard, G. Foster, P. Isabelle, Using Cognates to Align Sentences in Parallel Corpora, *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal (1992).

Smart Distance for Information Systems: The Concept

Yiming Ye, Prabir Nandi and Santhosh Kumaran[1]

Abstract—**We propose the concept of a "smart distance" for information systems and illustrate how to use it to interwoven various dynamic and heterogeneous elements of the system. With "smart distance" infrastructure operating within an information system, it can eliminate inefficiencies that are due to lags and latencies that exist in the traditional environment. On a larger scale, with "smart distance" operating across WWW, the Web can be more adaptive, useful, and popular.**

I. Background

THE pervasive connectivity of the Internet, coupled with an increasing distribution of organizations, are introducing profound changes in the way enterprises are set up and operated and are intensifying the forces within and across enterprises. To remain competitive under this environment, organizations need to move fast and to quickly adapt to business-induced changes. It must be able to sense the salient information, transform it into meaningful, quality business metrics, respond by driving the execution of business decisions into operational systems; and finally track the results against actions and expectations.

In parallel with this trend, there have been interesting developments in the fields of Intelligent Agents and Distributed Artificial Intelligence (DAI), notably in the concepts, theories and deployment of intelligent agents as a means of distributing computer-based problem solving expertise. Intelligent agents are well suited to the emerging character of the adaptive enterprise in which the distributed operations must be orchestrated into a synchronous flow.

There have been arguments or efforts to use agents or business objects to construct adaptive systems or enterprises [1] [4] [5] [8] [11] [12] [14]. For example, Hayes-Roth [8] studies the issue of how to build agents that function effectively in "adaptive intelligent systems" (AISs) that vary dynamically along dimensions like task requirements, different resources, contextual conditions, and performance criteria. She argues that an agent must adapt several key aspects of its behavior to dynamic situation such as its perceptual strategy, its control mode, its choices of reasoning tasks to perform, its choices of reasoning methods for performing those tasks, and its meta-control strategy for global coordination of all of its behavior. Despite various efforts in studying object oriented or agent oriented adaptive enterprises, as far as we are aware of, there are no working systems being widely used in practice. This is because that the state of the art of artificial intelligence has not reached to a stage such that an adaptive system can be operated effectively without human beings involvement.

In this paper, we propose the concept of a "smart distance" for complex enterprise systems and illustrate how this can be implemented by using Adaptive Document (ADoc) [12]. The idea is to wrap any business artifacts where "smart distance" is needed with a layer that enables this functionality. This layer is realized by extending an ADoc with relevant "smart distance" components. This will guarantee that the configurations and interactions of the elements of an enterprise system can be best placed, at any time and under any contextual environment. The concept of a "smart distance" comes from the observation that people like to adjust the "distances" among them when there are choices (for details, see [16]). The concept is further extended by including various elements in an enterprise system such as databases, business objects and intelligent agents, in addition to human beings. We define a "smart distance" in an organization as distances that are autonomously and adaptively adjusted based on contextual information with the goal that tasks can be best performed. Our approach can guarantee that the "distances" of various elements of an information system are adaptively placed to favor the task, at any time and under any contextual environment.

II. Smart Distance for Business Artifacts

Here we illustrate the concept of a "smart distance" for an enterprise system. As we have discussed before, this concept comes from the observation of human-human interaction. It is extended to agent supported collaborative work where intelligent agents are used to adjust the communication channels among people based on contextual information [16]. Here, we further extend the concept to an adaptive enterprise system to refer to the situation that different artifacts dynamically adjust their "distance" configurations such that the performance of the enterprise can be maximized. These artifacts can be users of the enterprise, business entity agents, and business objects, etc.

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA {yiming,prabir,sbk}@us.ibm.com

The "distance" refers to the various degrees of awareness, communications, and interactions among different artifacts.

The smart distance concept can be intuitively represented in Fig. 1. Smart distance between people refers to the situation where people intelligently adjust their distance based on various social contexts and preferences (Fig. 1(a)). Fig. 1(b) shows the concept of a "smart distance" between people and object. The person in the figure adjusts the distance between his eyes and the books under different contexts and vision conditions. Fig. 1(c) shows the natural distance, or the best distance between two artifacts under a contextual situation at a given time. Fig. 1(d) shows the real distance between two artifacts, which may not be the natural distance. Fig. 1(e) shows the situation that the two artifacts perform some autonomous actions such that their distance can be the same as their natural distance.

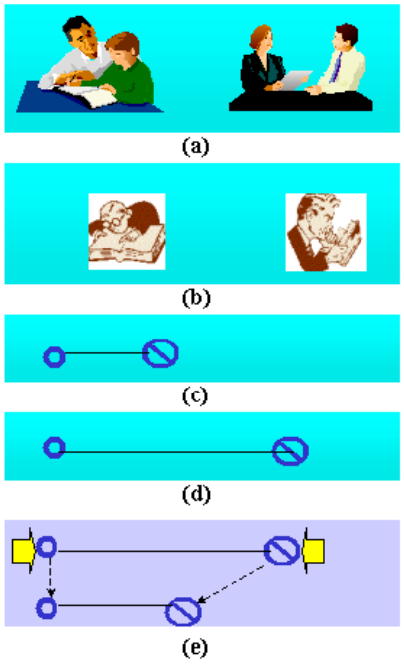


Fig. 1. Smart distance illustration.

The following is a formal definition for smart distance for an enterprise system.

- E is an Enterprise system
- $A = \{a_1, \cdots, a_n\}$ is the set of all the artifacts of this enterprise system. An artifact can be an object, an autonomous element, an intelligent agent, a database system, an information system, or a human being etc.
- For any two artifacts a_i and a_j, suppose that there are I_{ij} different kinds of interactions for a_i to interact with a_j. We call each kind of interaction a "channel". For example, people to people interaction through web can have video channel, audio channel, text channel.
- For a given channel c, there might be different degree of interactions; we used $|c| \in [0,1]$ to represent the degree of interaction. The bigger the value of $|c|$, the more intense the interaction for this channel. With $|c| = 1$ to represent the strongest interaction for this channel, and $|c| = 0$ to represent the weakest interaction through this channel. In most case, $|c| = 0$ means that this channel is closed. For example, the degree of interaction for a video channel from one person to another can be defined as the resolution and the update rates of the video transmission. The degree of interaction from a monitoring agent to a user can be defined as the frequency that the monitored data is sent to the user.
- The distance from a_i to a_j can thus be represented as a vector: $d_{ij} = \langle |c_1|, \cdots, |c_{I_{ij}}| \rangle$.
- The distance configurations (at time τ) for a given enterprise can thus be represented by a matrix:
$$D(\tau) = \begin{pmatrix} d_{11}(\tau), \cdots, d_{1n}(\tau) \\ \cdots \\ d_{n1}(\tau), \cdots, d_{nn}(\tau) \end{pmatrix}$$
- Under a given contextual/environmental condition at time τ, $\Omega(\tau)$ (it might contain many parameters), there exists a natural distance configuration (or called best placed distance configuration) $D_{natural}(\Omega(\tau))$. If the distance configurations of all the artifacts in the enterprise E is equal to this natural distance, that is, if $D(\tau) = D_{natural}(\Omega(\tau))$, then the performance is maximized.
- In most situations, $D(\tau) \neq D_{natural}(\Omega(\tau))$
- Smart distance for an enterprise system means that the artifacts in an enterprise act autonomously such that $\| D(\tau) - D_{natural}(\Omega(\tau)) \|$ is minimized. This is a challenging task.

Our goal is to construct the enterprise system such that all the artifacts will be woven by "smart distance". Since usually best distance configurations are very difficult to be pre-determined/pre-coded because of the complexity and the dynamics of the environment, we need to provide a way of adjusting these natural distances easily.

III. Example Scenario

We use the following example to explain how smart distances among different enterprise users can be implemented with Adaptive Document (ADoc) [12]. An ADoc is a business object that is implemented using Enterprise Java Bean. It has state machines inside the bean to

specify its states under various contextual situations. It also provides a convenient GUI for user interaction.

Suppose that a big retailer store has a virtual information system to process the supplier related issues. Suppose that there are a large number of distributed human agents associated with the retailer to handle the contract issues with various suppliers. Whenever a supplier has a product to offer, it will log onto the system and input supply related information. Then a human agent from the store will take over the issue and negotiate with the supplier in order to have the contract signed. Since there are a large number of human agents for the store, there is no way for the human agents to track each other's activity and to interact among each other. However, these may not be good to the retailer because appropriate interactions among human agents might help the retailer to do better deals under certain contextual situations.

For each user agent, we construct an ADoc to represent it. Within the ADoc, there is a business process template pool that provides all the different business processes that this user agent is associated with. For example, the left side of Fig. 2 shows a simple business process that a contract manager might be involved. First, a supplier logs onto the information system requesting to supply a product. Then, the contractor and the supplier will discuss about the contract and make a series of modifications. When the draft is finalized, it will be sent to the upper management chain for approval. The results will be that either the draft is approved or rejected.

Awareness among contract managers is needed during the

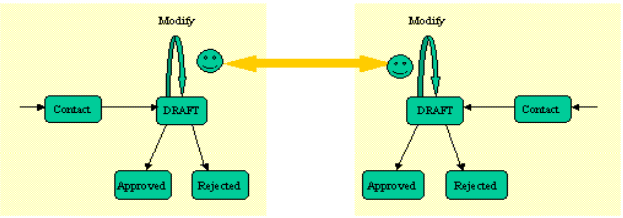


Fig. 2. Smart distance between contractors of the retailer store. The state machines represent the business flows of the contracting process.

"Draft" stage. Because it is possible that the two contracts in progress might be related to the same product. Thus, from the enterprise point of view, the two managers should be aware of each other's activities and interact with each other so as to get the best deal in the contracts (Fig. 2). This functionality is realized by extended ADOC (corresponding to the "Smart Distance" module of the BESA agent architecture) and by the "Smart Distance" directory of the enterprise server.

IV. Local Registry

Whenever a human agent logged onto the system, an ADoc is generated to represent the user. This user will then input through the ADoc UI to specify his awareness requirements, or in other words, his "smart distance" requirements under different contextual situations. This ADoc will track the user's activities and provide the "smart distance" functionality based on the user's registration. The local awareness registry will contain a Hash table. Each item of the table gives one awareness requirement that contains a list of information. Fig. 3 shows a specific example UML representation of the local registry for smart distance for the current example. It is an extension of ADOC [12]. The distance values in this example are: no awareness, Sametime [9] interaction, Phone interaction, Check document, and Meeting scheduling. Please note that Fig. 3 only gives a specific model of a local registration. In general, different systems may model the local registrations quite differently. The following list explains the details of the most important elements for a generic smart distance extension with respect to a single registration item.

- Awareness Condition for the current agent. This entry gives the condition to start the Smart Distance process, or in other words, the condition to start choosing distance values.
  - BPT_ID: Business Process Template ID. For example, the Template for Contract business process as illustrated in Fig. 2. In Fig. 3, it is represented by "My State Machine ID" of the "AW Condition Class".
  - State_ID: the state for the BPT_ID. For example, the "DRAFT" state in Fig. 2. In Fig. 3, this information is represented by "My StateID" of the "AW Condition Class".
- Condition on "when" to interact with "which" agents:
  - AE_Category: the category of the other agent to look for. In the above example, the category of the agent to look for is also a contract agent. In Fig. 3, this is represented by "HisAgentID" of the "AW Condition" class.
  - BPT_ID: the business process that the other agent is currently in. In Fig. 3, it is represented by "HisStateMachineID".
  - State_ID: the state of the BPT_ID the other agent is at. In Fig. 3, it is represented by "HisStateID".
- Actions: this gives the awareness actions list to be taken. These actions are distance values. For each action, the following information needs to be registered. There might be more than one action to be taken.
  - Awareness_ID. This gives the category of awareness to be requested. The list of awareness categories is listed in the global server that will be discussed in the next section. One example of the awareness is to provide a communication channel such as open a SameTime [9] window for interaction. Some awareness categories might be quite complex and may even contain the interaction choreography or conversational policies and protocols that specify a structured interaction. In Fig. 3, the awareness selections is given by the "AW ActionList" which provides all the awareness

requirements that need to be satisfied. Please note that when the Cardinality is "0", then the corresponding awareness channel will not be opened.

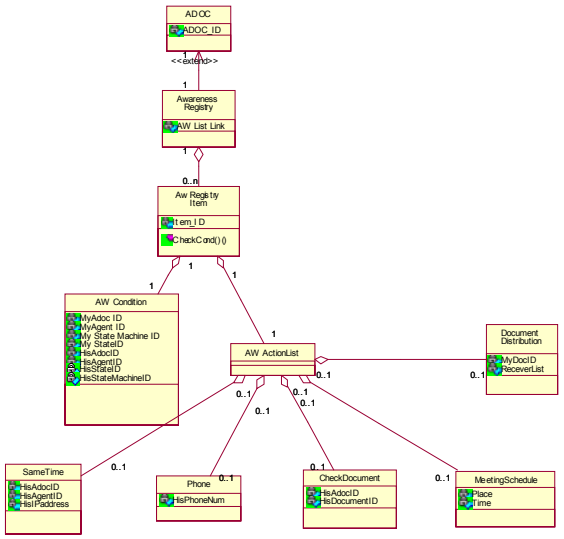
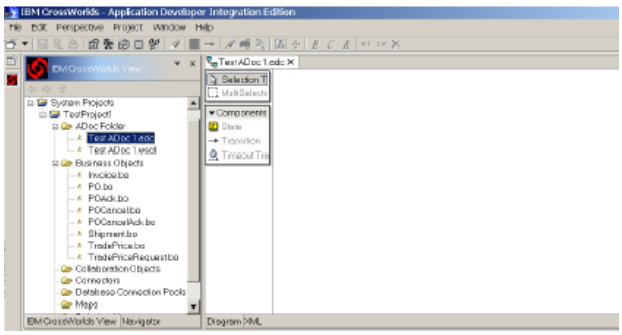
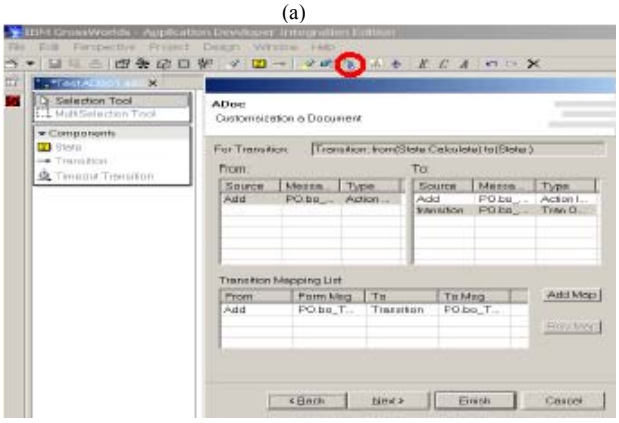


Fig. 3. UML representation of the local registry.

To implement the extension of ADoc to handle smart distance, we can either directly modify the code or use the ADoc builder that is currently being developed [17]. Fig. 4 shows the "state machine editor" and the customization Wizard of the ADoc builder.



(a)



(b)

Fig. 4. (a) ADoc state machine editor. (b) ADOC customization wizard.

V. Global Awareness Server

The awareness functionality during the runtime is realized by a Global "Smart Distance" Awareness Server. The local registration of an agent gives its user's preference for distances. However, sometimes the enterprise might want to control the awareness level among various agents from a global point of view. For example, one user might want to interact with his director at a certain state. However, the enterprise might not allow this interaction because the user should interact with his/her first line manager for the corresponding issue. Thus, based on the contextual condition, the Global Smart Distance Server can enable individual awareness requests and also disable individual awareness requests.

Whenever a user agent logs onto the system, an ADoc will be generated to represent the user. After the user input the Local registration information, it will be transferred to the Global Registry. The Global Awareness Server collects all the Local Registration information from all the user agents. During the run time, the Global Awareness Server tracks the activities and performs matchmaking constantly so as to enable awareness requests when conditions are met.

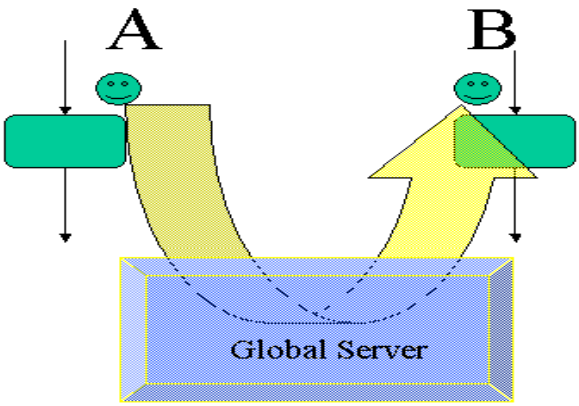


Fig. 5. Global server controls the degree on the channel from agent A to agent B during the run time.

Here is a simple example illustrating how the smart distance mechanism works during the run time. We assume the same store and the same virtual information system as described in Section 3. We also assume that the Global Awareness Server has a rule specifying that a communication channel will not be open if it only satisfies one agent's requests.

1. A supplier logs into the system, wants to supply tables to the store
2. John, a contract agent for the store, takes the job and signs in.
3. A contract ADoc is generated with ID_1, for the purpose of representing John.
   a. ADoc ID = 1
   b. John, as an agent, get an agent ID = A1
4. Through the ADoc User Interface, John inputs the following information into the Local Registration.

a. Condition
   i. John's Info:
      1. ID=A1
      2. State="Draft
      3. Object="Table" (the product from the supplier).
   ii. Other Agent's info (the information on other agents that adaptive communication might be needed):
      1. ADoc ID = x (x means that any ADoc will be considered).
      2. Agent ID = x (x means any agent will be considered)
      3. State="Draft"
      4. Object="Table"

b. Action
   i. Open sametime chat window for both agents if they are not in the same building
   ii. A phone call to schedule a face to face meeting if the agents are in the same building.

5. John's registration is propagated to the Global Registry. Global registry server will track John (A1)'s status and will check the condition info. whenever John's status is updated.
6. A supplier for "chair" logs into the system.
7. Mary takes the job and signs in.
8. An ADoc document ID_5 is generated and an Agent ID=A5 is generated for Mary.
9. Mary registered the local registry. She requests a sametime communication if there is another agent that is in the "Draft" state during a contract process and if the contract is related to either tables or chairs.
10. John and Mary are both at the Draft state. This is detected by the Global Awareness Server.
11. The communication channel for John and Mary does not open because only Mary's condition is satisfied (John wants interactions only when the other party is also doing a contract related to tables).
12. Another supplier for "table" logs onto the system.
13. Peter, who is at another site of the store, takes the job.
14. An ADOC is generated with ID_7 for Peter.
15. Peter registers the exactly same awareness condition as John.
16. Peter proceeds to the Draft state.
17. The Global Awareness Server responds to John's and Peter's requests and opens the sametime window for their communication.

VI. Discussions

In this paper, we propose the concept of a smart distance for complex enterprise systems and illustrate how to use ADoc to interwoven various business artifacts and to realize this functionality.

Our idea is to use ADoc to wrap any business artifacts that need "smart distance" functionality (Fig. 6(a)). These artifacts can be any elements of the system, ranging from

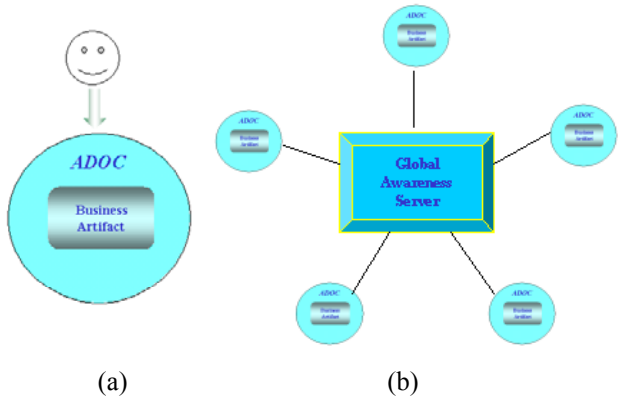


Fig. 6. (a) Use ADoc to wrap a business object in an enterprise. The UI of ADoc makes the change of distance requirement possible. (b) "Smart distance" is enabled with a Global Awareness Server. Please note that ADoc is only one way of realizing "smart distance". Many other approaches can be used. For example, we may use Enterprise Java Beans to implement the functionality.

business objects and databases to enterprise users. The UI functionality of the ADoc makes the adaptive change of distance requirements under dynamic, changing contextual situations possible. For ease of illustration, the scenario discussed in this paper is related to provide "smart distance" among users. However, the approach can be used to enable "smart distance" among various business objects within and across information systems. An information system with "smart distance" connecting its business artifacts can be more adaptive and on demand. On a larger scale, if we consider the World Wide Web as an ubiquitous information system, then providing smart distance among the huge number of artifacts of the WWW will make the Web even more useful, adaptive, and popular. Smart distance over the WWW can provide not only adaptive social networking among Web users, but also adaptive "social networking" among the huge number of Web artifacts. One approach to enable this is to build standards on "smart distance" over Web artifacts such that Web users can have a common language to specify their "smart distance" requirements that are understandable to programs or agents. We predict that this will be a reality in the future.

Acknowledgement

The authors would like to thank David Cohn for his support on the idea proposed in this paper and for his valuable suggestions on the possible application of this idea to business process monitoring. The authors would also like to thank Michel Desmarais for his valuable comments on the content and writing and Jiming Liu for his suggestions on the re-organization of this article.

REFERENCES

[1] J. Bigus, The Agent Building and Learning Environment, Proceedings of International Conference on Autonomous Agents, June 3-7, Barcelona, Spain, Pages 108-109, 2000.

[2] M.L. Brodie. The emperor's clothes are object oriented and distributed. In M.P. Papazoglou and G. Schlageter, editors, Cooperative Information Sys-tems: Trends and Directions, pages 15–48. Academic Press, 1998.

[3] Simon Cheng, Mathews Thomas, Santhosh Kumaran, Amaresh Rajasekharan, Fred Wu, Yiming Ye, Ying Huang. A Model-driven Approach for Item Synchronization and Uccnet Integration in Large E-Commerce Enterprise Systems, 5th International Conference on Enterprise Information Systems, Angers, France, 23-26, April 2003.

[4] P. Eeles and O. Sims. Building Business Objects. John Wiley & Sons, New York, 1998.

[5] B. Grosz, L. Hunsberger, and S. Kraus, Planning and Acting Together. AI Magazine, 20(4):23-34. 1999.

[6] James E. Hanson, Prabir Nandi, and David W. Levine, Conversation-enabled Web Services for Agents and e-business, Proceedings of the International Conference on Internet Computing (IC-02), CSREA Press, 2002, pp.791-796.

[7] J. Hendler, and A. Agrawala. Mission critical planning: AI on the MARUTI real-time operating system. In Proceedings of the Workshop on Innovative Approaches to Planning, Scheduling, and Control, 77-84, November, 1990.

[8] B. Hayes-Roth, An architecture for Adaptive Intelligent Systems, Artificial Intelligence Journal, 1993?

[9] http://www.lotus.com/products/lotussametime.nsf/wdocs/homepage

[10] http://www-3.ibm.com/software/integration/holosofx/1126announce.html

[11] Maamar, Z. and J. Sutherland. Toward Intelligent Business Objects: Focusing on techniques to enhance BOs that exhibit goal-oriented behaviors. Communi-cations of the ACM 43(10): 99-101.

[12] Prabir Nandi, Santhosh Kumaran, Terry Heath, Raja Das, Kumar Bhaskaran, ADoc-Oriented Programming, The 2003 International Symposium on Applications and the Internet (SAINT'2003), Jan 27-31, 2003, Orlando, Florida, USA.

[13] Papazoglou, M. Agent Oriented Technology in Support of E-Business: Enabling the Development of "Intelligent" Business Agents for Adaptive, ReusableSoftware Communications of the ACM, 44(4):71-77

[14] J. Sutherland, W.J. Heuvel (2002), Enterprise Application Integration Encounters Complex Adaptive Systems: A Business Object Perspective. Proceedings of the 35th Hawaii International Conference on System Sciences , Hawaii, USA.

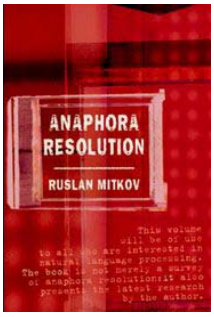
[15] Wu, L.S.Y., Hosking, J.R.M., and Doll, J.M., "Business Planning Under Uncertainty: Will We Attain Our Goal?", IBM Research Report RC 16120, Sep. 24, 1990

[16] Y. Ye, S. Boies, P. Huang, J. Tsotsos. (2001). Agents-Supported Adaptive Group Awareness: Samrt Distance and WWWaware. IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans, Vol. 31, No. 5. pp: 369-379.

[17] Jun Zhu and Zongwei Luo, ADoc Builder Manual Draft, 2002.

ANAPHORA RESOLUTION

By Ruslan Mitkov, Longman, 2002, ISBN: 0582325056



Reviewed by Nicolas Nicolov[1]

In the past two decades, as the amount of available information has been growing almost exponentially and data has become ever so plentiful, the gap between existing knowledge resources (in textual, audio and video form) and the ability of computer systems to extract that very knowledge has also been alarmingly widening. The dream of having a piece of wood becoming human, speak and *understand* (Pinocchio) has developed into the idea of having machines not only *speak*, but also autonomously *learn* - that is, acquire knowledge about the world from their language interactions with the environment (as S1mØne, a simulated computer agent, does in a Andrew Niccol's 2002 movie with Al Pacino). The latter aspect of understanding language has been an eagerly anticipated event by early AI researchers, which, by the way, didn't fail to be faced with at times bitter disappointments in those initial years; recently focus has shifted to more tractable concrete problems that allows us to make progress in building models that take advantage of enhanced Natural Language Processing (NLP) capabilities.

Anaphora describes the language phenomenon of referring to a previously mentioned *entity* (also called *object* or *event*); *anaphora resolution* is the process of finding that previous item. Consider the following clarifying example from a British World War II anti-raid leaflet:

"If an incendiary bomb drops next to you, don't loose your head. Put *it* in a bucket and cover *it* with sand."

If this raised eyebrows - don't worry - it is meant to. Indeed "it" could stand for (or *refer* to) either of the two objects mentioned before it, "bomb" and "head". The authors meant the former, but the rules of language have a tendency to bias readers to picking the latter. But then "head"s are not the usual things one puts in buckets and covers with sand. What anaphora resolution, when done correctly, enables us and systems to do, is to merge the previous information about an entity with the new information we encounter. Collecting dispersed pieces of information on an object ultimately builds a fuller picture about it; in technical parlance, systems can store the isolated pieces of knowledge in a knowledge base associated with the *same* object. And the more information we have in the knowledge base, the more new information can be automatically inferred (perhaps using the automated theorem proving technique of resolution as in Horn clause logic). So think of anaphora as the delicate balance between conciseness of communication and the ability of humans to understand each other.

A number of applications, Mitkov says, hinge on systems being able to do anaphora resolution right: machine translation, automatic abstracting, information extraction, question answering. NLP is the arena where the computer scientist meets the linguist; approaches to anaphora resolution require intricate understanding of language phenomena and making them operational requires solid computer science. In the book, thus, Mitkov is addressing a wide audience and illustrating concisely, yet thoroughly, the needed prerequisites. In "Einstein$_i$ felt he$_i$ was on the right track" *he* refers to Einstein (hence we put the same indices *i*). In this case we are lucky there is only one possible item *he* could point back to. Mitkov stresses that more often than not in normal circumstances (texts be them monologues or dialogues) there would be a number of items that could potentially be referred to; this ambiguity gives rise to the need for resolving the correct anaphor (previous item). Although as humans we are amazingly good at picking the right referent (the technical term for the previous item is *antecedent*), for machines this is by far not so straightforward. While the discussion so far might misleadingly suggest it is only pronouns that have the magic property of making us search our mental representations for matching items (pronominal anaphors), Mitkov quickly gives a comprehensive classification of anaphoric phenomena, including "invisible things" (*zero anaphor*) being able to magically refer back. The mystery is due to a peculiarity of Romance languages; in the Italian example, "Judy e' molto intelligente; si e' laureata alla Edinburgh University" ("Judy is very intelligent; [she] is a graduate of Edinburgh University"), Italian allows, actually expects, speakers to drop the pronoun "she" because the morphology of the (reflexive) verb "si e' laureata" [3rd person singular, feminine] makes it clear that Judy is the intended subject in the second sentence.[2]

Conversely languages also allow for pointing back to entities that haven't been mentioned explicitly: "As John was driving, a rabbit jumped on to the road and John slammed on the breaks." "the breaks" refers to the vehicle John is driv-

[1]IBM, nicolas@watson.ibm.com

[2]Incidentally, if you ever wondered why in Italian we can skip the pronouns when the [3rd person singular, present tense] verb provides such strong indications as what the pronoun could be and in English in the same situation we cannot - rest assured in Italian all other pronouns are dropped (hence also the name pro-drop languages). In English the verbs forms for present tense verbs which are not [3rd person singular] are the same and speakers of English would have a harder task of picking the right antecedent.

ing which was not mentioned explicitly.

Mitkov then goes on to describe declaratively what knowledge sources are evolved in the process of anaphora resolution. Consider "The children$_1$ had sweets$_2$. They$_?$ were deli _____." Substituting "delighted" and "delicious" for the last word yields two different antecedents for the pronoun "they", "the children" and "the sweets" respectively (linguists routinely use substitution tests to demonstrate certain constructions are possible and others are kind of odd). The example is a clear case where semantic knowledge about what entities can be delighted and what entities can be delicious helps in picking the right antecedent.

Returning to the applications that need or greatly benefit from anaphora resolution engines, machine translation can take advantage of them, using similar anaphoric expressions in the translation output if the languages are close (e.g., translating from Norwegian into Swedish). Automatic abstracting (summarization), information extraction and question answering all stress the need of being able to piece together knowledge about entities or events which is spread through the information source (not all facts about a person are stated at the point when they are first introduced in the text). Something that Mitkov does not mention (and traditionally not considered a part of anaphora resolution) is that anaphora plays a role in the language generation process. Getting it wrong, as in our World War II leaflet example, brings smiles to people's faces. And in order not to get in wrong technical documentation guidelines recognize the inherent ambiguity of possible antecedents and explicitly try to reduce the chance of the reader getting the antecedent wrong by avoiding certain anaphoric constructions. Incidentally, writing guidelines for lazy readers resort to the same technique - save the reader the effort of finding what you meant by telling him explicitly (perhaps risking a bit of repetition).

Anaphora was appreciated quickly enough as a stumbling block in furthering progress in NLP and a number of theoretical approaches and systems emerged in early 1980s to deal with it. Mitkov does empower the reader with succinct coverage of the theories and the knowledge-intensive techniques of the 80s. AI wasn't "situated" then and researchers would make assumptions about what (pre-)processing was available to them making for elegant theoretical frameworks but not resulting in systems that could easily be applied in practice. Mitkov contrasts this knowledge-intensive approach with later developments that impose fewer requirements on the depths of preprocessing. He calls these techniques knowledge-poor, and as one might guess, these are techniques that derive their "poor" knowledge from corpora. "The pressing need for the development of robust and inexpensive solutions to meet the demands of practical NLP systems encouraged many researchers to move away from extensive domain and linguistic knowledge and to embark instead upon knowledge-poor anaphora resolution strategies." (page 94). An additional factor that enabled less knowledge-intensive approaches to be explored was the availability of both common tools and corpora that permitted the use of machine learning techniques. And finally the field was viewed ripe enough that conferences included tracks on resolving anaphors - the Message Understanding Conferences (MUC6 & MUC7) gave considerable momentum to research in the area. More recently, the Automatic Content Extraction (ACE) evaluation also crucially includes resolving anaphors. Multilinguality is also a factor of concern and researchers are interested in domain- and language-independent techniques. Different languages, though, exhibit subtle differences in the kinds of anaphors they use and their distributions.

Mitkov covers a lot of ground and necessarily at various points needs to refer the reader to the original sources for greater details, though the description of the techniques allows for rational reconstruction of the original work. He does, however, change pace and presents as a comprehensive case study the approach and system he has been developing over the years (MARS). This is the place in the book where the practice of building anaphora resolution engines is fully revealed. The goal is to describe a fully automatic, knowledge-poor, multilingual system. Mitkov does notice a drop of performance when the system works on real output of pre-processing components which are not perfect and make errors; he suggests that previous research should be examined critically in view of many systems having been evaluated under the assumption that they had had access to perfect preprocessing of the input.

The proliferation of approaches and systems begs the question "How do I, as a natural language engineer, choose among alternative anaphora resolution engines?" Corpora with coreference links allow direct comparisons. Mitkov draws a distinction between evaluating an anaphora resolution algorithm and evaluating an anaphora resolution engine as a component of a larger system. For algorithms he presents precision and recall measures, performance measures, comparative evaluation tasks and component measures. For systems Mitkov presents an evaluation workbench where in a plug-and-play mode different engines can be substituted for and the change in performance characteristics observed.

Finally, Mitkov concludes by taking a step back and considering the accomplishments of research in the area of anaphora resolution so far (*Centering* theory about entities in the focus of the attention of the speaker and listener, *Discourse Representation Theory* and how discourse elements are accessed, wide array of systems using different levels of knowledge). He then considers present challenges and directions of future research. Researchers actively working in the area of anaphora resolution as well as graduate students should look here for ways to push the frontiers of science even further.

So are we really close to the moment when S1mØne can understand the questions posed to her without Al Pacino frantically pushing buttons to produce her response? Mitkov says we are 80% there but covering the remaining 20% will not be easy.

RELATED CONFERENCES & CALL FOR PAPERS

TCCI Sponsored Conferences

ICDM'03
The Third IEEE International Conference on Data Mining
Melbourne, Florida, USA
November 19-22, 2003
http://www.cs.uvm.edu/∼xwu/icdm-03.html
Submission Deadline: June 10, 2003

The 2003 IEEE International Conference on Data Mining (IEEE ICDM '03) provides a leading international forum for the sharing of original research results and practical development experiences among researchers and application developers from different data mining related areas such as machine learning, automated scientific discovery, statistics, pattern recognition, knowledge acquisition, soft computing, databases and data warehousing, data visualization, and knowledge-based systems. The conference seeks solutions to challenging problems facing the development of data mining systems, and shapes future directions of research by promoting high quality, novel and daring research findings. As an important part of the conference, the workshops program will focus on new research challenges and initiatives, and the tutorial program will cover emerging data mining technologies and the state-of-the-art of data mining developments.

Topics related to the design, analysis and implementation of data mining theory, systems and applications are of interest. These include, but are not limited to the following areas: foundations of data mining, data mining algorithms and methods in traditional areas (such as classification, clustering, probabilistic modeling, and association analysis), and in new areas, mining text and semi-structured data, and mining temporal, spatial and multimedia data, data and knowledge representation for data mining, complexity, efficiency, and scalability issues in data mining, data pre-processing, data reduction, feature selection and feature transformation, post-processing of data mining results, statistics and probability in large-scale data mining, soft computing (including neural networks, fuzzy logic, evolutionary computation, and rough sets) and uncertainty management for data mining, integration of data warehousing, OLAP and data mining, human-machine interaction and visualization in data mining, and visual data mining, high performance and distributed data mining, machine learning, pattern recognition and scientific discovery, quality assessment and interestingness metrics of data mining results, process-centric data mining and models of data mining process, security, privacy and social impact of data mining, and data mining applications in electronic commerce, bioinformatics, computer security, Web intelligence, intelligent learning database systems, finance, marketing, healthcare, telecommunications, and other fields.

WI 2003
The 2003 IEEE/WIC International Conference on Web Intelligence
Halifax, Canada
October 13-17, 2003
http://www.comp.hkbu.edu.hk/WI03/

Web Intelligence (WI) is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of Artificial Intelligence (AI) and advanced Information Technology (IT) on the next generation of Web-empowered products, systems, services, and activities. It is the key and the most urgent research field of IT in the era of World Wide Web and agent intelligence. The IEEE/WIC 2003 joint conferences are sponsored and organized by IEEE Computer Society Technical Committee on Computational Intelligence (TCCI) and by Web Intelligence Consortium (WIC), an international organization dedicated to the promotion of world-wide scientific research and industrial development in the era of Web and agent intelligence.

The technical issues to be addressed include, but not limited to: Intelligent Web-Based Business, Knowledge Networks and Management, Ubiquitous Computing and Social Intelligence, Intelligent Human-Web Interaction, Web Information Management, Web Information Retrieval, Web Agents, Web Mining and Farming, Emerging Web Technology.

IAT 2003
The 2003 IEEE/WIC International Conference on Intelligent Agent Technology
Halifax, Canada
October 13-17, 2003
http://www.comp.hkbu.edu.hk/IAT03/

The 2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT 2003) is a high-quality, high-impact agent conference, which is jointly held with the 2003 IEEE/WIC International Conference on Web Intelligence (WI 2003). The IEEE/WIC 2003 joint conferences are sponsored and organized by IEEE Computer Society Technical Committee on Computational Intelligence (TCCI) and by Web Intelligence Consortium (WIC), an international organization dedicated to the promotion of world-wide scientific research and industrial development in the era of Web and agent intelligence.

The upcoming meeting in this conference series follows the great success of IAT-99 held in Hong Kong in 1999 (http://www.comp.hkbu.edu.hk/IAT99/) and IAT-01 held in Maebashi City, Japan in 2001 (http://kis.maebashi-it.ac.jp/iat01/). The aim of IAT 2003 is to bring together researchers and practitioners from diverse fields, such as computer science, information technology, business, education, human factors, systems engineering, and robotics to (1) examine the design principles and performance characteristics of various approaches in intelligent agent technology, and (2) increase the cross-fertilization of ideas on the development of autonomous agents and multi-agent systems among different domains. The technical issues to be addressed include, but not limited to: Applications, Computational Models, Architecture, and Infrastructure, Autonomy-Oriented Computation (AOC) Paradigm, Learning and Self-Adapting Agents, Data and Knowledge Management Agents, and Distributed Intelligence.

Other Computational Intelligence Conferences

AAMAS'03
The Second International Joint Conference on Autonomous Agents and Multi-Agent Systems
Melbourne, Australia
July 14-18, 2003
http://www.aamas-conference.org

Agents are one of the most prominent and attractive technologies in computer science at the beginning of the new milennium. The technologies, methods, and theories of agents and multiagent systems are currently contributing to many diverse domains such as information retrieval, user interfaces, electronic commerce, robotics, computer mediated collaboration, computer games, education and training, ubiquitous computing, and social simulation. They not only are a very promising technology, but are also emerging as a new way of thinking, a conceptual paradigm for analyzing problems and for designing systems, for dealing with complexity, distribution, and interactivity, while providing a new perspective on computing and intelligence. The AAMAS conferences aim to bring together the world's researchers

active in this important, vibrant, and rapidly growing field.

The AAMAS conference series was initiated in 2002 as a merger of three highly respected individual conferences (ICMAS, AGENTS, and ATAL). The aim of the joint conference is to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multiagent systems.

IJCAI'03
The Eighteenth International Joint Conference on Artificial Intelligence
Acapulco, Mexico
August 9-15, 2003
http://www.ijcai-03.org/1024/index.html

The IJCAI-03 Program Committee invites submissions of full technical papers for IJCAI-03, to be held in Acapulco, Mexico, August 9-15, 2003. Submissions are invited on substantial, original, and previously unpublished research on all aspects of Artificial Intelligence. Refer to the web site for the list of related keywords.

ISMIS 2003
The Fourteenth International Symposium on Methodologies for Intelligent Systems
Maebashi TERRSA, Maebashi City, Japan
October 28-31, 2003
http://www.wi-lab.com/ismis03/

This Symposium is intended to attract individuals who are actively engaged both in theoretical and practical aspects of intelligent systems. The goal is to provide a platform for a useful exchange between theoreticians and practitioners, and to foster the cross-fertilization of ideas in the following areas: active media human-computer interaction, autonomic and evolutionary computation, intelligent agent technology, intelligent information retrieval, intelligent information systems, knowledge representation and integration, knowledge discovery and data mining, logic for artificial intelligence, soft computing, and Web intelligence. In addition, we solicit papers dealing with Applications of Intelligent Systems in complex/novel domains, e.g. human genome, global change, manufacturing, health care, etc.

evonet

UNIVERSITÀ DEGLI STUDI DI PARMA
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Call For Participation: EvoNet Summer School 2003 in Evolutionary Computing

University of Parma, Italy
24 - 31 August 2003
http://www.evonet.info/summerschool2003/

The 6th EvoNet Summer School is for everyone who wants to learn about evolutionary computing and how to apply it in real world problems. Organised by the EvoNet Network of Excellence and the University of Parma, it is especially aimed at PhD students, postdocs, researchers and practitioners, and repeats a winning formula of presenting challenging problems with appropriate tools and tutorials to solve them in practical sessions. The focus is on teamwork, collaboration and pooling resources with numbers limited to maximise learning and student interaction.

Highlights include:

- A selection of challenging real-life problems and easy-to-use tools
- Problem solving in small teams
- Introductory lectures for first-time users and advanced tutorials for more experienced researchers
- Practical, hand-on sessions to maximise learning
- Skill tutorials on conducting good research, writing papers and giving presentations
- Guidance from leading researchers in European evolutionary computing

Problems areas and the Senior Researchers presenting them will include:

- Exploring mechanisms to deal with problems embedded in dynamic environments - Ernesto Costa, University of Coimbra
- Evolution and analysis of neural robot controllers - Tom Ziemke, University of Skvde
- Automatic Concept Evolution - Terry Fogarty, South Bank University
- N-Player Iterated Prisoner's Dilemma Games - Xin Yao, University of Birmingham
- Fast Advanced Unconventional Genetic Programming - Riccardo Poli, University of Essex
- A 2D cutting problem from glass industry - Gunther Raidl, Vienna University of Technology

The EvoNet Summer School will be held in the historic town of Parma with its rich cultural and gastronomic traditions. Local Organiser: Stefano Cagnoni, Department of Computer Engineering, University of Parma.

Call For Papers and Special Issues Proposals:

The IEEE Computational Intelligence Bulletin (ISSN 1727-5997 & ISSN 1727-6004)

The IEEE Computational Intelligence Bulletin is the official publication of the Technical Committee on Computational Intelligence (TCCI) of the IEEE Computer Society. The Bulletin publishes **technical or survey articles** as well as **special thematic issues**. We welcome manuscripts that focus on original theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. Manuscripts should be submitted to the Associate Editors of respective sections.

If you plan to propose or organize a special issue in certain areas of general interest, please feel free to contact Dr. Jiming Liu, the Bulletin's Editor-in-Chief, via email at `jiming@comp.hkbu.edu.hk`.

Call For Papers: IEEE Intelligent Systems Special Issue

Mining the Web for Actionable Knowledge

Submissions due 1 Aug. 2003

Recently, there is much work on data mining on the Web to discover novel and useful knowledge about the Web and its users. Much of this knowledge can be consumed directly by computers rather than humans. Such actionable knowledge can be applied back to the Web for measurable performance improvement. For example, knowledge about user behavior and usage patterns can be used to build adaptive user interfaces and high-performance network systems. Web site structure provides important content and linkage information that can be used for Web searches. Web pages can be wrapped for building more convenient Web services. This special issue of IEEE Intelligent Systems will feature articles that address the problem of actionable data mining on the Web. We are particularly interested in papers that offer measurable gains in terms of well-defined performance criteria through Web data mining.

For this special issue, we invite original, high-quality submissions that address all aspects of Web mining for actionable knowledge. Submissions must address the issues of what knowledge is discovered and how such knowledge is applied to improve the performance of Web based systems. Topics of interest include but are not limited to

- Web information extraction and wrapping
- Web resource discovery and topic distillation
- Web search
- Web services
- Web mining for searching, querying, and crawling
- Web content personalization
- Adaptive Web sites
- Adaptive Web caching and prefetching

Special Issue Guest Editors

- Craig Knoblock, University of Southern California, Information Sciences Institute
- Xindong Wu, University of Vermont
- Qiang Yang, Hong Kong University of Science and Technology

Important Dates

- 1 Aug. 2003: Submissions due
- 5 Sept. 2003: Notification of acceptance
- 7 Nov. 2003: Final version submitted
- 9 Jan. 2004: Issue ships

Submission Guidelines

Submissions should be 3,000 to 7,500 words (counting a standard figure or table as 250 words) and should follow the magazines style and presentation guidelines (see `http://computer.org/intelligent/author.htm`). References should be limited to 10 citations. Send submissions to `qyang@cs.ust.hk`.

Call For Papers: Special Session on

Multi-agent Learning, Game-theoretic and Negotiation Agents in CIRAS03

http://ciras.nus.edu.sg/2003/mas_sim.pdf

DEADLINE: August 1, 2003

Special Issue in Journal

==========

Extended versions of selected high quality papers from this session will be invited to submit to a special issue on Learning Approaches for Negotiation Agents and Automated Negotiation in the International Journal of Intelligent Systems (Guest Editor: Professor Kwang M. Sim).

Page length: Max. 6 pages, 2-column IEEE Format.

Instructions for preparing manuscripts are provided at http://ciras.nus.edu.sg/2003/submission.html.

Submit all manuscripts in pdf format via e-mail to: Professor Kwang M. SIM, kmsim@ie.cuhk.edu.hk
http://www.ie.cuhk.edu.hk/people/ben.php

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398