THE IEEE Intelligent Informatics BULLETIN

IEEE COMPUTER SOCIETY

December 2011 Vol. 12 No. 1 (ISSN 1727–5997)

IEEE Computer Society
Technical Committee
on Intelligent Informatics

Profile

The FMS Cognitive Modeling Group at Carnegie Mellon University. . . . .*Christian Lebiere* 1

Conference Report

Crossroads in Constraint Programming . . . .*Marius Silaghi & Christian Bessiere* 4
Progress and Challenges in Product Configuration Systems Development . . . .*Dietmar Jannach* 6

Feature Articles

Influence Propagation in Social Networks: A Data Mining Perspective . . . . *Francesco Bonchi* 8
Mining a Data Reasoning Model for Personalized Text Classification. . . . .*Luepol Pipanmaekaporn & Yuefeng Li* 17
Analysing Effect of Database Grouping on Multi-Database Mining. . . . . *Animesh Adhikari, Lakhmi C. Jain & Sheela Ramanna* 25

Book Review

Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data . . . .*Russell Johnson* 33

Special Issue Review

Neurocomputing - Special Issue on Advances in Web Intelligence . . . . *Stefan Rüger, Vijay Raghavan, Irwin King & Jimmy Xiangji Huang* 35

Announcements

Related Conferences, Call For Papers/Participants. 37

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

IEEE Computer Society Technical Committee on Intelligent Informatics (TCII)

Executive Committee of the TCII:

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using biologically and linguistically motivated computational paradigms such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. If you are a member of the IEEE Computer Society, you may join the TCII without cost at http://computer.org/tcsignup/.

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once 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 TCII Executive Committee

2) Feature Articles

3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)

4) Book Reviews

5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics 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.

Editorial Board

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@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).*

The FMS Cognitive Modeling Group at Carnegie Mellon University

Scaling Cognitive Models to All Levels of Human Activity

I. Cognitive Architectures

Understanding the workings of human cognition remains a fundamental scientific challenge: while the basic building blocks of the brain are well known, as is their general organization, there is no agreement on how cognition emerges from their interaction. Unified theories of cognition consisting of invariant mechanisms and representations, implemented computationally as cognitive architectures, have been proposed as a way to organize the empirical findings and master the complexity of neural systems. If successful, they would also represent the most promising way of engineering artificial systems capable of general intelligence, a.k.a. Strong AI or AGI.

ACT-R is an integrated computational cognitive architecture resulting from decades of cumulative effort by an international community of cognitive researchers. It consists of a modular framework (see figure 1) with the following components: a) procedural and declarative memory modules, including both symbolic and subsymbolic (i.e., statistical) representation and learning mechanisms; b) perceptual and motor modules that incorporate many known human factors parameters and provide principled limitations on the interaction with an external environment; c) a constrained modular framework for incorporating additional factors such as fatigue and emotions that are not currently part of the architecture; and d) asynchronous interaction between modules that assemble small, sub-second cognitive steps into complex streams of cognition to accomplish high-level functionality.

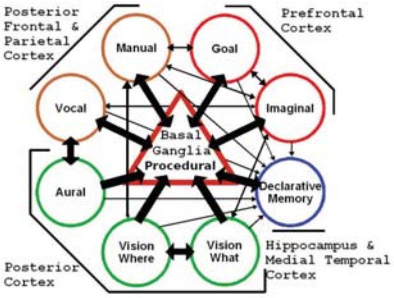
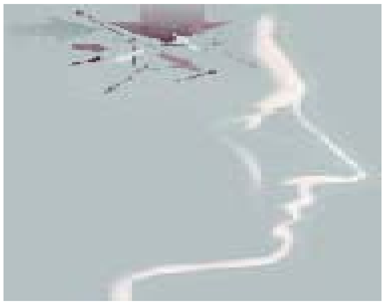


Fig. 1. Overview of the ACT-R Architecture, showing the different modules classified according to brain region.



Models build using the architecture can learn to perform complex dynamic tasks while interacting directly with the same environment as human users. ACT-R can account for all quantitative measures of human performance, from behavioral measures such as response time, percent of correct responses and eye movements, to fine-grained neural measurements such as EEG and fMRI. Hundred of cognitive models have been validated against experimental data, for tasks ranging from performing simple psychology experiments to controlling complex information systems such as air traffic control.

II. Research Agenda

ACT-R models have adopted a particular level of abstraction, usually interact with abstract simulation systems, and represent the cognition of a single agent with definite goals. Our current research aims at lifting those limitations by relating to other levels of description of our cognitive system, in particular the neural level, by grounding ACT-R within physical systems such as robots, and by extending the single-model paradigm to multiple models interacting in cognitive networks.

A. Cognitive Neuroscience

The primary challenge in cognitive science is to construct bridges between levels of description similar to those erected in the physical sciences. For instance, when understanding how people gather and process information in decision-making, there are various levels at which these decision processes can be analyzed. Using Marr's tri-level approach to information processing, at the computational level, sensemaking accounts of the decision-making process are formulated in terms of large-scale situational awareness and descriptive knowledge structures.

In contrast, at the implementation level, neurally inspired models provide accounts of brain regions that implement the processes described by sensemaking. Bridging these levels is the algorithmic level, which is captured by functional models such as ACT-R bridging sensemaking goals with neural architecture.

The difficulty lies in translating high-level qualitative sensemaking descriptions to the language of neurological mechanism. We use the ACT-R cognitive architecture to better understand the cognitive processes underlying sensemaking and provide a bridge to the implementational level.

Our most recent project in that direction is a collaborative effort devoted to understanding cognitive biases in the context of geospatial intelligence analysis by building neural models of artificial tasks, with sensemaking as the theoretical framework for information processing.

High-fidelity models of information processing at the neural level are created by our partners at UC Boulder using the Emergent architecture, a framework for developing neural networks that utilize the Leabra learning algorithm.

The role of our group in this project is to create functional cognitive models of the tasks in ACT-R. Doing so serves two purposes. The first is to provide a bridge between sensemaking theory and neural theory. The second is to prototype models of tasks in order to assist the Emergent modelers by providing functional constraints on the neural models. That enterprise is scientifically grounded by the architectural commitments shared by ACT-R and Leabra despite their distinct levels of description.

Mechanisms that were prototyped in ACT-R include perceptual path planning, anchoring and adjustment bias in learning, within-participant probability matching, confirmation bias in information selection and resource allocation in intelligence analysis.

B. Cognitive Robotics

A major issue limiting the application of cognitive architectures as general intelligent systems is their lack of physical grounding. Real-world intelligence is grounded in our experience in the physical world. To achieve comparable integration and autonomy in the world such as on robotic platforms, we are extending our collaboration with the Leabra architecture. Specifically, we are developing a hybrid framework called SAL ("Synthesis of ACT-R and Leabra") to leverage the strengths of both architectures. The current instantiation of the SAL framework uses Leabra for perception, leveraging neural adaptivity, while ACT-R handles tasks such as decision-making that rely on symbolic control.

One research goal for the hybrid architecture is to create a system capable of unsupervised learning. To accomplish this, a visual stimulus is presented to Leabra, which then passes the resulting high-level neural representation of the stimulus to ACT-R. ACT-R then determines via declarative memory recalls whether the pattern presented is sufficiently similar to one already known; if not, ACT-R generates an arbitrary label for the stimulus, and issues a command to Leabra to train the visual system on the given stimulus using the ACT-R-generated label. This process is then iterated until a stable labeling of the stimulus on both sides of the hybrid framework is achieved.

Eventually, the hybrid model will learn not just to develop its own categories, but to learn the affordances of objects and to manipulate them. The ongoing development is focused on completing the loop between perception and cognition, and integrating motor functions into the framework. The ultimate goal is a fully autonomous system that learns to perceive its world and act upon it without any assistance.

In parallel, we have also started exploring integrating ACT-R with algorithmic approaches to perception. We use a state-of-the-art object detection algorithm to detect pedestrians walking along a sidewalk. Information from perception is then fed into an ACT-R model that tracks these pedestrians as they make their way along the sidewalk. The model uses these tracks to categorize (and, in some cases, predict) pedestrian behavior as either normal or suspicious. Categorization/prediction works on the basis of the model checking the calculated tracks for the presence of certain features together with other contextual information (such as the presence or absence of certain objects/people). We are currently working on detecting the relevant features automatically from the perceptual data by observing differences between expected and actual behavior.

A long-term goal of this research is to generalize the use of expectations by the model to drive cognitive behavior. In the expectation-driven approach, a model would generate an expected outcome for any proposed action. When this action is performed, the model compares the expected outcome with the real outcome. The result of this comparison is used in selecting future actions. We believe that such an expectation-driven approach has implications all along the cognitive and rational bands in Newell's Time Scale of Human Cognition, from attention and learning to task selection and planning.

C. Visual Intelligence

Beyond basic perceptual grounding, an important focus of our research is on developing models of visual intelligence, namely the capability to learn representations of actions from visual inputs, reason over them, and eventually generate a meaningful description in natural language.

Far from being a bare summation over object features detected in the environment, visual intelligence has to be conceived as a complex phenomenon where perception combines with high-level cognitive processes and conceptual knowledge. A key distinction between this research and the state of the art in computer vision is that the latter exclusively deals with detection and classification of a wide range of objects (nouns in the description of a scene, e.g. "person", "car", "ball", etc.) and features (e.g., position, orientation, shape, direction of motion, etc.) while research in visual intelligence operates at a more composite and unifying level: the goal is to identify the perceptual and cognitive underpinnings for recognizing and reasoning about actions (denoted by verbs in natural language), focusing on the roles played by objects in a scene (e.g., "a person hits a car with a ball"). Human understanding of events results from intertwined perceptual, cognitive and inferential processes: reproducing this capability at the machine level requires a comprehensive infrastructure where optical invariants, low-level perceptual mechanisms and high-level cognitive processes couple with knowledge representations.

In this framework we are working on integrating the ACT-R architecture (cognitive level) with computational ontologies (representational level), aiming at building a hybrid knowledge system where cognitive mechanisms are combined with representational contents. In particular, computational ontologies specify the meaning of (conceptual) representations of the world, encoding the semantics of those knowledge contents in a computational model.

The challenge of this research is to develop a full-fledged system where reasoning capabilities are cognitively grounded and driven by mechanisms of perception, memory and control. In this context, the purely technological problem of augmenting the ACT-R cognitive architecture with suitable ontologies of actions and automatic reasoners depends on the scientific problem of understanding the intertwined dynamics of reasoning, knowledge formation and perceptual mechanisms: in this sense, developing cognitively-inspired visual intelligence systems can also be seen, in a broader context, as a step forward in cognitive science and artificial intelligence.

D. Economics and game theory

Traditionally, cognitive models, like participants in psychology experiments, are given specific goals to accomplish. In the real world, however, people are confronted with ill-defined tasks, conflicting incentives, and open-ended learning opportunities.

To embody models with the human ability to set goals and define their own task, we are working on a project aimed at understanding adversarial cognition and motivation. We use strategic games inspired from the field of game theory and enhanced to afford human experimentation. For instance, we are currently developing an experimental paradigm called IPD^3 – Intergroup Prisoner's Dilemma with Intragroup Power Dynamics and Individual Power Drive – that provides a useable interface between humans and computers to play a series of nested repeated games. We have designed the game so that it is solidly grounded in state-of-the-art game-theoretic and socio-cognitive research. This design retains features that are advantageous for experimental purposes (e.g., binary choice, matrix format, computational tractability) while adding features that increase ecological validity (e.g., multiple players, social structure, asymmetries, conflicting motives, and stochastic behavior). We use this paradigm to collect human data in carefully designed experiments. In parallel, we develop cognitive models of human behavior in strategic games. For example, in one empirical study we found that human participants are perfect reciprocators: negative emotions triggered by unreciprocated attempts at cooperation bias subsequent decision-making toward defection. However, a cognitive model of the same game learned that sustained cooperation (even when occasionally unreciprocated) was more effective. The cognitive model did not show the retaliation bias because it lacked affective processing. We are working on developing an affective module for the ACT-R cognitive architecture that will allow us to model the full range of human cognition and emotion in strategic games. Our cognitive modeling approach complements the traditional equilibrium analyses in predicting the effect of game modifications. In repeated games, almost any outcome can be an equilibrium. Game simulations with validated cognitive models as players can be used to narrow down the set of possible equilibria to a limited number of cognitively plausible outcomes, generate specific predictions about human behavior in these games, and provide more believable synthetic characters for games and training.

E. Network science

Beyond the interactions of small groups of individuals lies the emergence of civilization from the collective cognitive acts of large-scale societies. Among the most fascinating abilities of human beings is their propensity to verbalize, communicate and adopt ideas within a vast network of social contacts. Human cognitive capabilities are uniquely suited to communication, and they are crucial to the intelligence emerging from human communities. The cognitive and psycholinguistic mechanisms underlying language comprehension and production are still poorly understood. While recent studies paint a picture of how memory and contextualization help humans comprehend a dialogue partner's ideas and individual language, we do not understand whether human memory has evolved to support teamwork and social cognition. In addition to communication, other tasks such as decision-making under uncertainty are crucially enhanced by teamwork, despite the fast and frugal heuristics or performance bounds found in individual human actors.

Cognitive modeling and network simulation techniques have allowed a recent growth in interest for the interaction of cognitive mechanisms with the social environment. Cognitive architectures can characterize the bounded human abilities to recall information, which is key in explaining the interactions of humans in a network.

In recent work, we have shown how individuals adapt their linguistic expressions quickly to their interaction partners, and new communicative conventions soon spread through a network of connected entities, which may be cognitive models of humans or simple software agents.

Cognitive modeling frameworks, validated and refined through careful experimentation, as well as computational tools now allow the larger-scale simulation of human societies and the uptake of existing language resources (corpora) in the quest for the architecture of the human language faculty. We have developed a networked experimentation platform called the Geogame to facilitate large-scale data collection. Datasets collected in real-life situations let us test cognitive and psycholinguistic models. Once validated, they will make better predictions and cover broad ranges of human behaviour. This combination of broad coverage and large-scale simulation requires new computational tools, new methodologies, new datasets and new experimental designs. With ACT-UP, our lab has developed a rapid-prototyping scalable implementation of the ACT-R cognitive architecture, allowing large-scale simulations of underspecified models.

III. Acknowledgements

The research described in this profile was supported by grants from AFOSR, DTRA, ARL, DARPA and IARPA to Dr. Christian Lebiere. All views expressed are solely those of the authors and not of the funding agencies.

Contact Information
The FMS Group:
Dir.: Christian Lebiere - cl@cmu.edu
Ion Juvina
Unmesh Kurup
Alessandro Oltramari
David Reitter
Matthew Rutledge-Taylor
Robert Thomson
Jerry Vinokurov
Websites:
http://fms.psy.cmu.edu
http://act-r.psy.cmu.edu

Crossroads in Constraint Programming

Marius Silaghi, Florida Tech, USA, marius.silaghi@fit.edu
Christian Bessiere, University of Montpelier, France, bessiere@lirmm.fr

I. Constraint Programming

Constraint programming is a declarative programming paradigm exploiting techniques stemming from research on combinatorial problems in computer vision, and robot planning. The paradigm comes close to the dream that users only need to simply state a problem and the computer will solve it, as underlined in the seminal 1996 article *In Pursuit of the Holy Grail* by Eugene Freuder. The idea is that programming should be possible by simply stating the problem using a set of constraints. Using the traits of these constraints, appropriate techniques are then automatically selected and applied for solving the problem. Constraints are ubiquitous, and the requirement to satisfy them can be modeled within the framework of the *constraint satisfaction problem* (CSP). The constraints of a CSP are specified as relations on a set of variables. The choice of these variables and formulation of the constraints turns out to be essential for the efficiency of the obtained program. Many optimization problems can also be addressed with various extensions of constraint programming. Constraints whose satisfaction is optional are called *soft constraints*. They can be associated with a function that quantifies the desire for their satisfaction.

The brute force approach to combinatorial problems is usually considered to be one of either the *chronological backtracking* or the *generate and test* method. The research into constraint programming has started with work on local reasoning. Local reasoning combines a subset of the known constraints to infer new constraints. Therefore it is also known as *constraint propagation*. When a local inconsistency is inferred, the reduction in the size of the search space (Cartesian product of domains for variables) is potentially exponential. Propagation is desirable when the overhead is polynomial and promises exponential speed-up. The propagation process is also referred to as *local consistency enforcement* since new constraints illustrate clearer what values are allowed. In general, the generation of unary constraints, i.e., removing values from domains, which has lower overhead, has been more successful in solvers, specially when repeatedly applied on subproblems during backtracking. Techniques where all applied operations result only in redundant constraints or in the splitting of the search space guarantee that no solution is lost. Those techniques guaranteeing to find a solution whenever a solution exists are said to be *complete*. Most past research has focused on complete algorithms, and this is being identified as something that may change in the near future.

Fig. 1. Freuderfest location

If one views the variables of a CSP as nodes and the constraints relating them as arcs (or hyper-arcs), the obtained *graph structure* captures essential information about the problem. It was shown by Freuder that problems whose graph structure is a tree can be solved in linear time. Other properties of the graph structure were found for which polynomial time algorithms exist. Problems that do not originally present such structures can be split or otherwise processed to reduce them to the desirable structures. Constraints that involve a large number of variables, notably those that involve a number of variables dependent on the size of the problem, cannot be processed straightforwardly. Special propagation techniques have been developed for many types of such *global constraints*, hundreds of them being available in catalogs set up by researchers in the constraint programming community. It was observed that the structure of the CSP graphs as well as the internal structure of particular constraints presents symmetries and search on such symmetric parts is redundant, the results being directly transferable. Significant recent effort was placed on detecting and exploiting *symmetries*. The community is also exploring *distributed CSPs*, namely where some of the constraints are secrets of participants who share a desire to find values that satisfy all their constraints. The applications of CSPs have raised other research topics such as *constraint extraction* from examples or from text, and *constraint elicitation* from users. Another problem is how to provide *explanations* about which subset of the constraints could be changed to transform an insolvable problem into one that has solutions. Several centers for research into constraint programming concentrate researchers, the largest being the Cork Constraint Computation Center (4C).

II. Freuderfest and CP 2011

The 17th International Conference on Principles and Practice of Constraint Programming was held in September 2011 in Perugia, Italy. Freuderfest, a special workshop dedicated to the retirement of Eugene Freuder, the founder and director of the Constraint Computation Center in Cork was held the day before the conference in the historical building of the administration of Perugia. As it was highlighted there by Francesca Rossi, Eugene has worked with 111 collaborators on 112 articles. Some of his closest collaborators were invited to speak about the impact of his work on their research. The workshop was opened with an introduction by Barry O'Sullivan, the new director of 4C. Eugene himself used his speech to argue that the CP research community should debate more on the relevance of completeness and encouraged the intensification of research into incomplete techniques. This is a significant turn from his original focus on complete backtrack-free search and tree structures, interchangeability and symmetries. The other speakers focused on his seminal works in each of these fields, in which he is known to have launched and structured the basic ideas.

Eight collocated workshops were scheduled during the first day of the conference. Among them were the *11th Workshop on Soft Constraints* (Soft), the *8th Workshop on Local Search* (LSCS), *11th Workshop on Symmetry* (SymCon), and *10th Workshop on Modeling and Reformulation* (ModRef). There were also new workshops on *CSP based Bioinformatics* (WCB), *Component Configuration* (Lococo), *Parallel Methods* (PMCS), and on the *MiniZinc* Modeling language. The ModRef workshop organized a very popular panel on the available extensions to the common CSP representations. The discussions addressed the difficulties of extensions with Boolean Clauses, LP, MDDs and Neural Networks.

The main conference was started with an invited talk by Jean-Charles Régin on common pitfalls to avoid when solving problems with CP. His conclusion was that when CP does not work, most likely it is due to the choice of a bad representation. The Best Student Paper Award was offered to Marie Pelleau for her work on representing constraints in continuous domains using two Cartesian systems of coordinates rotated at 45 degrees. The solution intersects boxes found along those two different coordinate systems, significantly improving the approximation of complex shapes. Meinoff Sellman from IBM gave a tutorial on solver portfolios. He recommended SATzilla and CP Hydra as the best portfolio solvers for SAT and CSPs.

The Best Paper Award was offered to Georg Gottlob for proving an old conjecture about the NP-hardness of Minimal Constraint Networks. He shows that even when a constraint problem is reduced to its minimal configuration, namely where each pair of values for any two variables that is not forbidden by any constraint appears in at least one solution, finding a solution is a hard problem.

In a premiere for CP, the invited talk by Patrick Prosser with the occasion of receiving the Research Excellence Award from the Association of Constraint Programming (ACP) was delivered using a much appreciated YouTube video. The ACP Doctoral Research Award was offered to Standa Živný, who gave a talk on the complexity and expressive power of valued CSPs. He showed that many tractable constraint languages can be obtained by composing constraints from existing languages. The doctoral students tutorial by Laurent Michel focused on the importance of careful design for experiments, factoring out the noise and caching issues with sufficiently large problem sets.

Laurent Perron from Google described a successful application of constraint programming within the most important Google servers, namely for taking caching decisions and decisions concerning the routing of user traffic. The constraint programming technique has saved 1 millisecond which translates into savings of millions of dollars given the scale of the corresponding operations at Google. Google proposed the ROADEF challenge which stresses the need for a good solution fast (no optimality and no completeness required). The challenge has deadlines in each August and February. The Best Application Paper Award was offered for a paper coauthored by Venkatesh Ramamoorthy for successfully using CSPs to design highly nonlinear cryptographic functions. Such functions are the most sensible part of modern ciphers, and resistance to known attacks requires the satisfaction of a set of constraints. No solution is known to satisfy all these constraints, but some of them are soft and solutions have been found within certain thresholds. A few of these constraints are global and their decomposition, when done without preserving completeness, enabled to find functions that better satisfy the soft constraints when compared to previously found ones.

The conference was closed with a celebration of Gene Freuder's retirement, under the form of a panel on the future of constraint programming. Eugene stressed the importance of selecting lofty goals, launching buzzwords like: Problem solving Web, Electronic Embodiment, Computational Precognition and Business Constraints. Visible applications are in healthcare, analytics, energy, sustainability, and humanitarian operations. The researchers should balance the just-in-case redundant constraint generation tendency that challenges memory requirements for the just-in-time approach that spends resources only when needed. Approaching the lofty goals should be achieved also by broadening the scope of the CP conference which should join all research where the constraint is relevant, no longer focusing only on algorithms based on backtracking.

All the presentation slides were made available on website of the program of the conference: dmi.unipg.it/cp2011.

Fig. 2. Group picture at CP2011

Christian Bessiere is research professor at CNRS at Montpelier, France. His research has focused on constraint programming. He has published articles on all aspects of constraint programming and in particular on constraint propagation.

Marius Silaghi is assistant professor in Computer Sciences at Florida Tech and chair of the Human Decision Support Systems Laboratory. His main research interests lie in the application of artificial intelligence and cryptology to support decision making. He has authored articles on distributed constraint reasoning and techniques involving cryptology and artificial intelligence.

Progress and Challenges in Product Configuration Systems Development

DIETMAR JANNACH, TU DORTMUND, GERMANY, IJCAI 2011 CONFIGURATION WORKSHOP CO-CHAIR
DIETMAR.JANNACH@TU-DORTMUND.DE

I. CONFIGURATION SYSTEMS

Mass customization and configure-to-order production are modern business strategies in which the goods and services offered by a company are tailored and individualized according to the specific needs and preferences of the customer. The classical examples of such configurable products include cars, personal computers, and various types of comparably complex technical devices; today, however, you can also create your custom muesli, skateboard, or travel package over the Web.

Mass customization and configure-to-order are based on the idea that the final customer product can be assembled and configured based on a predefined set of components, to keep the individualization costs and the corresponding sales prices low. This contrasts with other strategies such as engineer-to-order, in which customer-individual goods are manufactured,

Fig. 1. Barcelona impressions

Product configuration systems (configurators) are software applications that play a central role in such business approaches, in particular because in most cases the individual components cannot be assembled to the final customer product in an arbitrary way. Typically, the configuration process is governed by a set of constraints. In the car domain, such a constraint could for example be that one certain type of engine is not available with an automatic gear box. The reason for the existence of a constraint can both be technical (e.g., because of mechanical incompatibilities) or marketing-related (e.g., based on pricing strategies).

The tasks of a configuration system usually include the interactive acquisition of the user preferences, checking for constraint violations, the automatic completion of partial configurations, the generation of sales quotes (in case the configurator is used as a front-end sales tool) or a bill-of-materials (if it is used for back-office automation).

Beside the problem of the seamless integration of configurator solutions into a company's existing software infrastructure or ERP system, the development of the above-mentioned core configurator functionalities is not a trivial task for various reasons. One of the core problems is that of *knowledge representation*, that is, the question of how to encode the given constraints in the system. Note that the corresponding knowledge bases can be huge and comprise hundreds or even thousands of rules. Beside that, they are also subject to frequent changes in particular in technical domains. Another issue is that of *computational complexity*. In the end, configurations can consist of thousands of components, for example, in the telecommunications domain. Finally, there is also the aspect of *user interaction*. Note that interactive elicitation of the requirements is not a one-shot process, but often requires multiple interaction loops, e.g., for situations in which constraints are violated and the user has to revise some of the choices.

Due to their inherent complexity, configuration problems have always been in the focus of researchers in different areas of Computer Science, and in particular of Artificial Intelligence (AI) researchers, for the last 30 years. In fact, one of the earliest successfully applied rule-based expert systems was a configurator for computer systems.

Since these early rule-based years, progress was made with respect to various aspects of the development of configuration systems. Regarding modeling and knowledge representation, a number of languages were used to represent the configuration knowledge. Some of the approaches used first order logic or its decidable subsets such as constraints, description logic, or answer set programming. Others are based on graphical modeling languages. Reasoning was based on logical inference, case-based reasoning, different constraint satisfaction schemes, resource-based reasoning as well as interactive propose-and-revise approaches. With respect to user interaction, proposals have, for example, been made toward the personalization of the user interface, the dynamic pre-selection of input values or strategies to recover from situations when the user requirements cannot be satisfied. Beside these core issues in configurator development, researchers also put forward ideas toward intelligent debugging support for the knowledge bases, algorithms for distributed configuration problem solving as well as approaches addressing the problem of re-configuring an existing system.

In general, research in configuration systems is application-oriented and focusing on the development of generalizable solutions for practical problems in industry. In fact, we can observe that several ideas developed in the research community in the last decade made their way into industrial practice and that configuration is among those fields where AI technology has been successfully integrated into software products and applied in commercial settings.

II. The Configuration Workshop at IJCAI’11

The International Joint Conference on Artificial Intelligence (IJCAI), which took place in Barcelona in July 2011, featured a one day workshop on configuration. After a AAAI Fall symposium in 1996, this year’s workshop was already the thirteenth in a series of successful configuration workshops held at major international AI events such as IJCAI, AAAI or ECAI since its first edition in 1999.

Since the inception of the workshop series, strong and continuing interest from industry can be observed including tool providers such as ILOG (now IBM) or Tacton AB, ERP vendors like SAP and Oracle as well as large corporations (like Siemens or ABB) but also smaller configurator companies.

This year’s workshop featured two invited speakers and a scientific program consisting of four long paper and three short paper presentations. In the opening keynote talk, Fabrizio Salvador from the Instituto de Empresa Business School in Madrid looked at configuration systems from the business perspective. In particular, he emphasized the importance of the effective management of information on feasible product configurations to achieve higher responsiveness towards the customer. As an additional factor, he discussed not yet fully tapped potentials of learning from past configurations.

Learning from past configurations – although with a different purpose – was also the main idea of the paper “Incremental prediction of configurator input values based on association rules” presented in the subsequent technical sessions. In this work, the goal was to make the interaction with a configuration tool more convenient for the end user by dynamically pre-selecting appropriate input values and thereby reducing the interaction efforts.

The main focus of this year’s workshop, however, was on modeling. The presented works at the workshop included (1) a new graphical modeling approach which covers not only the configurable artefact but also the production process; (2) an integrated method to model the variability in heterogenous product families based on different *views*; (3) a best-practice modeling guideline for knowledge engineers involved in configuration development; and (4) a new modeling language, which can be used not only for configuration modeling but also for re-configuration and simulation purposes.

The second key topic of the 2011 workshop was on knowledge representation and reasoning (KRR). In the second keynote, Gerhard Friedrich from the Alpen-Adria University Klagenfurt, Austria, first reviewed key milestones with respect to KRR in the configuration domain. He then illustrated current challenges and solutions with the help of hard real-world configuration problem and presented results of a comparison of various KRR frameworks including Answer Set Programming (ASP). The results demonstrated the significant progress that has been made in the area in the last few years so that these KRR frameworks can nowadays be applied to solve practically relevant problems.

Answer Set Programming was also the basis for a new reconfiguration presented in another technical paper in the workshop. In this work, the authors demonstrated how the configuration and reconfiguration problem can be encoded in the ASP framework and present experimental results which indicate the feasibility of the approach for practical applications.

An alternative problem encoding was finally presented in the paper “Enumeration of valid partial configurations”, where the authors show how incremental SAT solvers can be used for online computation of partial configurations. These partial configurations are then used to suggest possible assignments, thus reducing the information load and improving the quality of decisions made by a user while configuring a product.

Overall, the workshop highlighted that significant advances have been made in the configuration area over the last years with respect to configuration modeling, knowledge representation and reasoning, and in particular that modern AI technology is more and more on the way to be usable nearly “off-the-shelf” for practical applications.

Despite these advances, many opportunities for future research remain in the configuration domain. For example, it would be interesting to see how the developed techniques can be applied beyond classical hardware configuration problems, e.g., for software and service configuration or model transformation. In addition, also different questions of knowledge acquisition and how to better integrate data from existing sources such as ERP or product data management systems with the configurator are still open. Finally, user interface issues were historically slightly underrepresented in this research community. Given that more and more products for the end-customer can be individualized over the Web, more focus has to be put on techniques for building adequate configurator user interfaces including aspects of “intelligent” customer guidance or 3D-visualization.

The papers of the IJCAI 2011 Workshop on Configuration can be downloaded at http://ceur-ws.org/Vol-755.

Fig. 2. Barcelona impressions

Dietmar Jannach is a professor in Computer Science at TU Dortmund, Germany and chair of the e-Services Research Group. His main research interests lie in the application of artificial intelligence and knowledge-based systems technology to real-world problems in particular in e-business environments. He has authored numerous papers on intelligent sales support systems such as recommender systems or product configurators. Dietmar Jannach was also one of the co-founders of ConfigWorks GmbH, a company focusing on next-generation interactive recommendation and advisory systems. He was a co-chair and organizer of the configuration workshop and the workshop on “Intelligent Techniques for Web Personalization” at IJCAI’11 and also gave a tutorial on Recommender Systems.

Influence Propagation in Social Networks: A Data Mining Perspective

Francesco Bonchi*

***Abstract*—With the success of online social networks and microblogs such as Facebook, Flickr and Twitter, the phenomenon of influence exerted by users of such platforms on other users, and how it propagates in the network, has recently attracted the interest of computer scientists, information technologists, and marketing specialists. One of the key problems in this area is the identification of influential users, by targeting whom certain desirable marketing outcomes can be achieved. In this article we take a data mining perspective and we discuss what (and how) can be learned from the available traces of past propagations. While doing this we provide a brief overview of some recent progresses in this area and discuss some open problems.**

By no means this article must be intended as an exhaustive survey: it is instead (admittedly) a rather biased and personal perspective of the author on the topic of influence propagation in social networks.

***Index Terms*—Social Networks, Social Influence, Viral Marketing, Influence Maximization.**

I. On social influence and viral marketing

The study of the spread of influence through a social network has a long history in the social sciences. The first investigations focused on the adoption of medical [1] and agricultural innovations [2]. Later marketing researchers have investigated the "word-of-mouth" diffusion process for *viral marketing* applications [3], [4], [5], [6].

The basic assumption is that when users see their social contacts performing an action they may decide to perform the action themselves. In truth, when users perform an action, they may have any one of a number of reasons for doing so: they may have heard of it outside of the online social network and may have decided it is worthwhile; the action may be very popular (e.g., buying an iPhone 4S may be such an action); or they may be genuinely influenced by seeing their social contacts perform that action [7]. The literature on these topics in social sciences is wide, and reviewing it is beyond the scope of this article.

The idea behind viral marketing is that by targeting the most influential users in the network we can activate a chain-reaction of influence driven by word-of-mouth, in such a way that with a very small marketing cost we can actually reach a very large portion of the network. Selecting these key users in a wide graph is an interesting learning task that has received a great deal of attention in the last years (for surveys see [8] and Chapter 19 of [9]).

Other applications include personalized recommendations [10], [11] and feed ranking in social networks [12], [13]. Besides, patterns of influence can be taken as a sign of user trust and exploited for computing trust propagation [14], [15], [16], [17] in large networks and in P2P systems. Analyzing the spread of influence in social networks is also useful to understand how information propagates, and more in general it is related to the fields of epidemics and innovation adoption. With the explosion of microblogging platforms, such as Twitter, the analysis of influence and information propagation in these social media is gaining further popularity [18], [19], [20], [21].

Many of the applications mentioned above essentially assume that social influence exists as a real phenomenon. However several authors have challenged the fact that, regardless the existence of correlation between users behavior with their social context [22], this can be really credited to social influence. Even in the cases where some social influence can be observed, it is not always clear whether this can really propagate and drive viral cascades.

Watts challenges the very notion of influential users that are often assumed in viral marketing papers [23], [24], [25], [19]. Other researchers have focussed on the important problem of distinguishing real social influence from *homophily* and other external factors [26], [27], [28], [29]. Homophily is a term coined by sociologists in the 1950s to explain the tendency of individuals to associate and bond with similar others. This is usually expressed by the famous adage "birds of a feather flock together". Homophily assumes *selection*, i.e., the fact that it is the similarity between users to breed connections [27].

Anagnostopoulos *et al.* [26] develop techniques (e.g., *shuffle test* and *edge-reversal* test) to separate influence from correlation, showing that in Flickr, while there is substantial social correlation in tagging behavior, such correlation cannot be attributed to influence.

However other researchers have instead found evidence of social influence. Some popular (and somehow controversial [30]) findings are due to Christakis and Fowler [31] that report effects of social influence over the spread of obesity (and smoking, alcohol consumption, and other unhealthy – yet pleasant – habits). Crandall *et al.* [27] also propose a framework to analyze the interactions between social influence and homophily. Their empirical analysis over Wikipedia editors social network and LiveJournal blogspace confirms that there exists a feedback effect between users similarity and social influence, and that combining features based on social ties and similarity is more predictive of future behavior than either social influence or similarity features alone, showing that both social influence and one's own interests are drivers of future

*This article summarizes, extends, and complements the keynote that the author gave at WI/IAT2011 conference, whose slides are available at: `www.francescobonchi.com/wi2011.pdf`

F. Bonchi is with Yahoo! Research, Barcelona, Spain.
E-mail: bonchi@yahoo-inc.com

behavior and that they operate in relatively independent ways.

Cha *et al.* [32] present a data analysis of how picture popularity is distributed across the Flickr social network, and characterize the role played by social links in information propagation. Their analysis provides empirical evidence that the social links are the dominant method of information propagation, accounting for more than 50% of the spread of favorite-marked pictures. Moreover, they show that information spreading is limited to individuals who are within close proximity of the uploaders, and that spreading takes a long time at each hop, oppositely to the common expectations about the quick and wide spread of word-of-mouth effect.

Leskovec *et al.* show patterns of influence by studying person-to-person recommendation for purchasing books and videos, finding conditions under which such recommendations are successful [33], [34]. Hill *et al.* [35], analyze the adoption of a new telecommunications service and show that it is possible to predict with a certain confidence whether customers will sign up for a new calling plan once one of their phone contacts does the same.

These are just few examples among many studies reporting some evidence of social influence. In this article we do not aim at providing an exhaustive survey, nor we dare entering the debate on the existence of social influence at the philosophical/sociological level. We do not even discuss further how to distinguish between social influence, homophily and other factors, although we agree that it is an interesting research problem. Instead, we prefer to take an algorithmic and data mining perspective, focussing on available data and on developing learning frameworks for social influence analysis.

Once sociologists had to infer and reconstruct social networks by tracking people relations in the real world. This is obviously a challenging and costly task, even to produce moderately sized social networks. Fortunately nowadays, thanks to the success of online social networks, we can collect very large graphs of explicitly declared social relations. Moreover, and maybe more importantly, we can collect information about the users of these online social networks performing some actions (e.g., post messages, pictures, or videos, buy, comment, link, rate, share, like, retweet) and the time at which such actions are performed. Therefore we can track real propagations in social networks. If we observe in the data user v performing an action a at time t, and user u, which is a “friend” of v, performing the same action shortly after, say at time $t + \Delta$, then we can think that action a propagated from v to u. If we observe this happening frequently enough, for many different actions, then we can safely conclude that user v is indeed exerting some influence on u.

In the rest of this article we will focus on this kind of data, i.e., a database of past propagations in a social network. We will emphasize that when analyzing social influence, it is important to consider this data and not only the structure of the social graph. Moreover, as this database of propagations might be potentially huge, we will highlight the need for devising clever algorithms that, by exploiting some incrementality property, can perform the needed computation with as few scans of the database as possible.

II. Influence Maximization

Suppose we are given a social network, that is a graph whose nodes are users and links represent social relations among the users. Suppose we are also given the estimates of reciprocal influence between individuals connected in the network, and suppose that we want to push a new product in the market. The mining problem of *influence maximization* is the following: given such a network with influence estimates, how should one select the set of initial users so that they eventually influence the largest number of users in the social network. This problem has received a good deal of attention by the data mining research community in the last decade.

The first to consider the propagation of influence and the problem of identification of influential users by a data mining perspective are Domingos and Richardson [36], [37]. They model the problem by means of *Markov random fields* and provide heuristics for choosing the users to target. In particular, the marketing objective function to maximize is the global expected lift in profit, that is, intuitively, the difference between the expected profit obtained by employing a marketing strategy and the expected profit obtained using no strategy at all [38]. A Markov random field is an undirected graphical model representing the joint distribution over a set of random variables, where vertices are variables, and edges represent dependencies between variables. It is adopted in the context of influence propagation by modelling only the final state of the network at convergence as one large global set of interdependent random variables.

Kempe *et al.* [39] tackle roughly the same problem as a problem in discrete optimization, obtaining provable approximation guarantees in several preexisting models coming from mathematical sociology. In particular their work focuses on two fundamental propagation models, named *Linear Threshold Model* (LT) and *Independent Cascade Model* (IC). In both these models, at a given timestamp, each node is either active (an adopter of the innovation, or a customer which already purchased the product) or inactive, and each node’s tendency to become active increases monotonically as more of its neighbors become active. An active node never becomes inactive again. Time unfolds deterministically in discrete steps. As time unfolds, more and more of neighbors of an inactive node u become active, eventually making u become active, and u’s decision may in turn trigger further decisions by nodes to which u is connected.

In the IC model, when a node v first becomes active, say at time t, it is considered contagious. It has one chance of influencing each inactive neighbor u with probability $p_{v,u}$, independently of the history thus far. If the tentative succeeds, u becomes active at time $t + 1$. The probability $p_{v,u}$, that can be considered as the strength of the influence of v over u.

In the LT model, each node u is influenced by each neighbor v according to a weight $p_{v,u}$, such that the sum of incoming weights to u is no more than 1. Each node u chooses a threshold θ_u uniformly at random from $[0, 1]$. At any timestamp t, if the total weight from the active neighbors of an inactive node u is at least θ_u, then u becomes active at timestamp $t + 1$.

In both the models, the process repeats until no new node becomes active. Given a propagation model m (e.g., IC or LT) and an initial seed set $S \subseteq V$, the expected number of active nodes at the end of the process is the *expected (influence) spread* of S, denoted by $\sigma_m(S)$. Then the *influence maximization problem* is defined as follows: given a directed and edge-weighted social graph $G = (V, E, p)$, a propagation model m, and a number $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that $\sigma_m(S)$ is maximum.

Under both the IC and LT propagation models, this problem is $\mathbf{NP}$-hard [39]. Kempe *et al.*, however, showed that the function $\sigma_m(S)$ is *monotone* and *submodular*. Monotonicity says as the set of activated nodes grows, the likelihood of a node getting activated should not decrease. In other words, $S \subseteq T$ implies $\sigma_m(S) \leq \sigma_m(T)$. Submodularity intuitively says that the probability for an active node to activate some inactive node u does not increase if more nodes have already attempted to activate u (u is, so to say, more "marketing saturated"). This is also called *"the law of diminishing returns"*. More precisely, $\sigma_m(S \cup \{w\}) - \sigma_m(S) \geq \sigma_m(T \cup \{w\}) - \sigma_m(T)$ whenever $S \subseteq T$.

Thanks to these two properties we can have a simple greedy algorithm (see Algorithm 1), which provides an approximation guarantee. In fact, for any monotone submodular function f with $f(\emptyset) = 0$, the problem of finding a set S of size k such that $f(S)$ is maximum, can be approximated to within a factor of $(1 - 1/e)$ by the greedy algorithm, as shown in an old result by Nemhauser *et al.* [40]. This result carries over to the influence maximization problem [39], meaning that the seed set we produce by means of Algorithm 1 is guaranteed to have an expected spread $> 63\%$ of the expected spread of the optimal seed set.

Although simple, Algorithm 1 is computationally prohibitive. The complex step of the greedy algorithm is in line 3, where we select the node that provides the largest marginal gain $\sigma_m(S \cup \{v\}) - \sigma_m(S)$ with respect to the expected spread of the current seed set S. Indeed, computing the expected spread of given set of nodes is $\#\mathbf{P}$-hard under both the IC model [41], [13] and the LT model [42]. In their paper, Kempe *et al.* run Monte Carlo (MC) simulations of the propagation model for sufficiently many times to obtain an accurate estimate of the expected spread. In particular, they show that for any $\phi > 0$, there is a $\delta > 0$ such that by using $(1 + \delta)$-approximate values of the expected spread, we obtain a $(1 - 1/e - \phi)$-approximation for the influence maximization problem. However, running many propagation simulations (Kempe *et al.* report $10,000$ trials for each estimation in their experiments) is practically unfeasible on very large real-world social networks. Therefore, following [39] many researchers have focussed on developing methods for improving the efficiency and scalability of influence maximization algorithms, as discussed next.

Leskovec *et al.* [43] study the propagation problem by a different perspective namely *outbreak detection*: how to select nodes in a network in order to detect as quickly as possible the spread of a virus? They present a general methodology for near optimal sensor placement in these and related problems. They also prove that the influence maximization problem of [39] is a special case of their more general problem definition. By exploiting submodularity they develop an efficient algorithm based on a "lazy-forward" optimization in selecting new seeds, achieving near optimal placements, while being 700 times faster than the simple greedy algorithm.

Algorithm 1 Greedy alg. for influence maximization [39]

```
Require: $G, k, \sigma_m$
Ensure: seed set $S$
1: $S \leftarrow \emptyset$
2: while $|S| < k$ do
3:     $u \leftarrow \arg\max_{w \in V \setminus S}(\sigma_m(S \cup \{w\}) - \sigma_m(S))$;
4:     $S \leftarrow S \cup \{u\}$
```

Regardless of this big improvement over the basic greedy algorithm, their method still face serious scalability problems as shown in [44]. In that paper, Chen *et al.* improve the efficiency of the greedy algorithm and propose new degree discount heuristics that produce influence spread close to that of the greedy algorithm but much more efficiently.

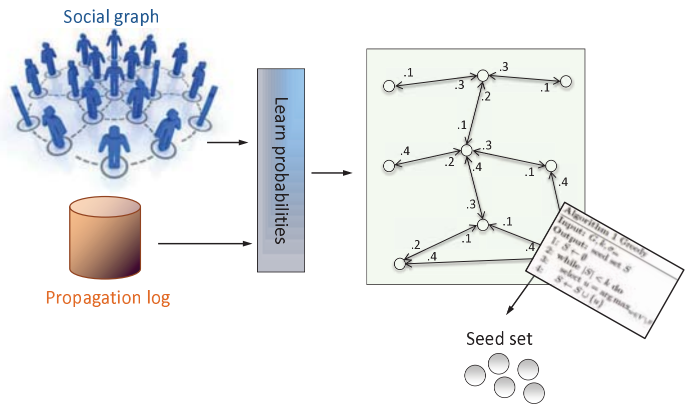
In their following work Chen *et al.* [41] propose scalable heuristics to estimate coverage of a set under the IC model by considering Maximum Influence Paths (MIP). A MIP between a pair of nodes (v, u) is the path with the maximum propagation probability from v to u. The idea is to restrict the influence propagation through the MIPs. Based on this, the authors propose two models: *maximum influence arborescence* (MIA) model and its extension, the *prefix excluding MIA* (PMIA) model.

Very recently, Chen *et al.* [42] proposed a scalable heuristic for the LT model. They observe that, while computing the expected spread (or coverage) is $\#\mathbf{P}$-hard in general graphs, it can be computed in linear time in DAGs (directed acyclic graphs). They exploit this property by constructing local DAGs (LDAG) for every node in the graph. A LDAG for user u contains the nodes that have significant influence over u (more than a given threshold θ). Based on this idea, they propose a heuristic called LDAG which provides close approximation to Algorithm 1 and is highly scalable.

III. Propagation traces

In most of the literature on influence maximization (as the set of papers discussed above), the directed link-weighted social graph is assumed as input to the problem. Probably due to the difficulties in finding real propagation traces, researchers have simply given for granted that we can learn the links probabilities (or weights) from some available past propagation data, without addressing how to actually do that (with the exception of few articles described in the next section). This way they have been able to just focus on developing algorithms for the problem which takes the already-weighted graph as input.

However, in order to run experiments, the edge influence weights/probabilities are needed. Thus researches have often assumed some trivial model of links probabilities for their experiments. For instance, for the IC model often experiments are conducted assuming *uniform* link probabilities (e.g., all



```
Algorithm 1 Greedy
Input: $G, k, \sigma_m$
Output: seed set $S$
1: $S \leftarrow \emptyset$
2: while $|S| < k$ do
3:     select $u = \arg\max_{w \in V \setminus S}$
4:     $S \leftarrow S \cup \{u\}$
```

Fig. 1. The standard influence maximization process.

links have probability $p = 0.01$), or the *trivalency* (*TV*) model where link probabilities are selected uniformly at random from the set $\{0.1, 0.01, 0.001\}$, or assuming the *weighted cascade* (*WC*) model, that is $p(u, v) = 1/d_v$ where d_v represent the in-degree of v (see e.g., [39], [41]).

These experiments usually are aimed at showing that a newly proposed heuristic select a seed set S much more efficiently than Algorithm 1, without losing too much in terms of expected spread achieved $\sigma_m(S)$.

In a recent paper Goyal *et al.* [45] have compared the different outcomes of the greedy Algorithm 1 under the IC model, when adopting different ways of assigning probabilities. In particular, they have compared the trivial models discussed above with influence probabilities learned from past propagation traces. This is done by means of two experiments on real-world datasets.

In the first experiment the overlap of the seed sets extracted under the different settings is measured. In the second experiment, the log of past propagations is divided in training and test set, where the training set is used for learning the probabilities. Then for each propagation in the test set, the set of users that are the first to participate in the propagation among their friends, i.e., the set of "initiators" of the action, is considered as the seed set, and the actual spread, i.e., the size of the propagation in the test set, is what the various methods have to predict.

The outcome of this experimentation is that: (i) the seed sets extracted under different probabilities settings are very different (with empty or very small intersection) , and (ii) the method based on learned probabilities outperforms the trivial methods of assigning probabilities in terms of accuracy in predicting the spread. The conclusion is hence that it is extremely important to exploit available past propagation traces to learn the probabilities.

In Figure 1, we summarize the standard process followed in influence maximization making explicit the phase of learning the link probabilities. The process starts with the (unweighted) social graph and a log of past action propagations that say when each user performed an action. The log is used to estimate influence probabilities among the nodes. This produces the directed link-weighted graph which is then given as input to the greedy algorithm to produce the seed set using MC simulations.

We can consider the propagation log to be a relational table with schema $(user\_ID, action\_ID, time)$. We say that an action propagates from node u to node v whenever u and v are socially linked (have an edge in the social graph), and u performs the action before v. In this case we can also assume that u contributes in influencing v to perform that action. From this perspective, an action propagation can be seen as a flow, i.e., a directed subgraph, over the underlying social network. It is worth noting, that such a flow is a DAG: it is directed, each node can have zero or more parents, and cycles are impossible due to the time constraint. Therefore, another way to consider the propagation log is as a database (a set) of DAGs, where each DAG is an instance of the social graph.

In the rest of this article we will always consider the same input consisting of two pieces: (1) the social graph, and (2) the log of past propagations. We will see how different problems and approaches can be defined based on this input.

IV. Learning the influence probabilities

Saito *et al.* [46] were the first to study how to learn the probabilities for the IC model from a set of past propagations. They neatly formalize the likelihood maximization problem and then apply Expectation Maximization (EM) to solve it.

However, their theoretical formulation has some limitations when it comes to practice. One main issue is that they assume as input propagations that have the same shape as they were generated by the IC model itself. This means that an input propagation trace is a sequence of sets of users $D_0, \ldots, D_n$, corresponding to the sets of users activated in the corresponding discrete time steps of the IC propagation. Moreover for each node $u \in D_i$ it must exists a neighbor v of u such that $v \in D_{i-1}$. This is obviously not the case in real-world propagation traces, and some pre-processing is needed to close this gap between the model and the real data (as discussed in [47], [45]).

Another practical limitation of the EM-based method is discussed by Goyal *et al.* [45]. Empirically they found that the seed nodes picked by the greedy algorithm – with the IC model and probabilities learned with the EM-based method [46] – are all nodes which perform a very small number of actions, often just one action, and should not be considered as high influential nodes. For instance, Goyal *et al.* [45] report that in one experiment the first seed selected is a node that in the propagation traces appears only once, i.e., it performs only one action. But this action propagates to 20 of its neighbors. As a result, the EM-based method ends up assigning probability 1.0 to the edges from that node to all its 20 neighbors, making it a high influence node, so much influential that it results being picked as the first seed by the greedy algorithm. Obviously, in reality, such node cannot be considered as a highly influential node since its influence is not statistically significant.

Finally, another practical limit of the EM-based method is its scalability, as it needs to update the influence probability associated to each edge in each iteration.

Goyal *et al.* also studied the problem of learning influence probabilities [48], but under a different model, i.e., an instance of the *General Threshold Model* (or the equivalent *General Cascade Model* [39]). They extended this model by making influence probabilities decay with time. Indeed it has been observed by various researchers in various domains and on real data, that the probability of influence propagation decays exponentially on time. This means that if u is going to re-do an action (e.g., re-tweet a post) of v, this is likely going to happen shortly after v has performed the action, or never.

Goyal *et al.* [48] propose three classes of influence probabilities models. The first class of models assumes the influence probabilities are static and do not change with time. The second class of models assumes they are continuous functions of time. In the experiments it turns out that time-aware models are by far more accurate, but they are very expensive to learn on large data sets, because they are not incremental. Thus, the authors propose an approximation, known as Discrete Time models, where the joint influence probabilities can be computed incrementally and thus efficiently.

Their results give evidence that Discrete Time models are as accurate as continuous time ones, while being order of magnitude faster to compute, thus representing a good trade-off between accuracy and efficiency.

As the propagation log might be potentially huge, Goyal *et al.* pay particular attention in minimizing the number of scans of the propagations needed. In particular, they devise algorithms that can learn all the models in no more than two scans.

In that work, factors such as the *influenceability* of a specific user, or how influence-driven is a certain action are also investigated.

Finally, the authors show that their methods can also be used to predict *whether* a user will perform an action and *when* with high accuracy, and the precision is higher for user which have an high influenceability score.

V. Direct mining approaches

So far we have followed the standard approach to the influence maximization problem as depicted in Figure 1. First use a log of past propagations to learn edge-wise influence probability, then recombine these probabilities together by means of a MC simulation, in order to estimate the expected spread of a set of nodes.

Recently new approaches emerged trying to mine directly the two pieces of input (the social graph and the propagation log) in order to build a model of the influence spread of a set of nodes, avoiding the approach based on influence probability learning and MC simulation.

Goyal *et al.* [45] take a different perspective on the definition of the expected spread $\sigma_m(S)$, which is the objective function of the influence maximization problem. Note that both the IC and LT models discussed previously are probabilistic in nature. In the IC model, coin flips decide whether an active node will succeed in activating its peers. In the LT model it is the node threshold chosen uniformly at random, together with the influence weights of active neighbors, that decides whether a node becomes active.

Under both models, we can think of a propagation trace as a *possible world*, i.e., a possible outcome of a set of probabilistic choices. Given a propagation model and a directed and edge-weighted social graph $G = (V, E, p)$, let $\mathbb{G}$ denote the set of all possible worlds. Independently of the model m chosen, the expected spread $\sigma_m(S)$ can be written as:

$$\sigma_m(S) = \sum_{X \in \mathbb{G}} Pr[X] \cdot \sigma_m^X(S) \tag{1}$$

where $\sigma_m^X(S)$ is the number of nodes reachable from S in the possible world X. The number of possible worlds is clearly exponential, thus the standard approach (MC simulations) is to sample a possible world $X \in \mathbb{G}$, compute $\sigma_m^X(S)$, and repeat until the number of sampled worlds is large enough.

We now rewrite Eq. (1), obtaining a different perspective. Let $path(S, u)$ be an indicator random variable that is 1 if there exists a directed path from the set S to u and 0 otherwise. Moreover let $path_X(S, u)$ denote the value of the random variable in a possible world $X \in \mathbb{G}$. Then we have:

$$\sigma_m^X(S) = \sum_{u \in V} path_X(S, u) \tag{2}$$

Substituting in (1) and rearranging the terms we have:

$$\sigma_m(S) = \sum_{u \in V} \sum_{X \in \mathbb{G}} Pr[X]\, path_X(S, u) \tag{3}$$

The value of a random variable averaged over all possible worlds is, by definition, its expectation. Moreover the expectation of an indicator random variable is simply the probability of the positive event.

$$\sigma_m(S) = \sum_{u \in V} E[path(S, u)] = \sum_{u \in V} Pr[path(S, u) = 1] \tag{4}$$

That is, the expected spread of a set S is the sum over each node $u \in V$, of the probability of the node u getting activated given that S is the initial seed set.

While the standard approach samples possible worlds from the perspective of Eq. (1), Goyal *et al.* [45] observe that real propagation traces are similar to possible worlds, except they are *"real available worlds"*. Thus they approach the computation of influence spread from the perspective of Eq. (4), i.e., estimate directly $Pr[path(S, u) = 1]$ using the propagation traces available in the propagation log.

In order to estimate $Pr[path(S, u) = 1]$ using available propagation traces, it is natural to interpret such quantity as the fraction of the actions initiated by S that propagated to u, given that S is the seed set. More precisely, we could estimate this probability as

$$\frac{|\{a \in \mathcal{A} | initiate(a, S)\, \&\, \exists t : (u, a, t) \in \mathbb{L}\}|}{|\{a \in \mathcal{A} | initiate(a, S)\}|}$$

where $\mathbb{L}$ denotes the propagation log, and $initiate(a, S)$ is true iff S is precisely the set of initiators of action a. Unfortunately, this approach suffers from a *sparsity issue* which is intrinsic to the influence maximization problem.

Consider for instance a node x which is a very influential user for half of the network, and another node y which is a very influential user for the other half of the network. Their union $\{x, y\}$ is likely to be a very good seed set, but we can not estimate its spread by using the fraction of the actions containing $\{x, y\}$, because we might not have any propagation in the data with $\{x, y\}$ as the actual seed set.

Summarizing, if we need to estimate $Pr[path(S, u) = 1]$ for any set S and node u, we will need an enormous number of propagation traces corresponding to various combinations, where each trace has as its initiator set precisely the required node set S. It is clearly impractical to find a real-world action log where this can be realized (unless somebody sets up a large scale human-based experiment, where many propagations are started with the desired seed sets). It should be noted that this sparsity issue, is also the reason why it is impractical to compare two different influence maximization methods on the basis of a ground truth.

To overcome this obstacle, the authors propose a "u-centric" perspective to the estimation of $Pr[path(S, u) = 1]$: they scan the propagation log and each time they observe u performing an action they distribute "credits" to the possible influencers of a node u, retracing backwards the propagation network. This model is named *"credit distribution"* model.

Another direct mining approach, although totally different from the credit distribution model, and *not* aimed at solving the influence maximization problem was proposed by Goyal *et al.* few years ago in [49], [50]. In these papers they propose a framework based on the discovery of *frequent pattern of influence*, by mining the social graph and the propagation log. The goal is to identify the "leaders" and their "tribes" of followers in a social network.

Inspired by frequent pattern mining and association rules mining [51], Goyal *et al.*, define the notion of leadership based on how frequently a user exhibits influential behavior. In particular a user u is considered *leader* w.r.t. an action a provided u performed a and within a chosen time bound after u performed a, a sufficient number of other users performed a. Moreover these other users must be reachable from u thus capturing the role social ties may have played. If a user is found to act as a leader for sufficiently many actions, then it is considered a leader.

A stronger notion of leadership might be based on requiring that w.r.t. each of a class of actions of interest, the set of influenced users are the same. To distinguish from the notion of leader above, Goyal *et al.* refer to this notion as *tribe leader*, meaning the user leads a fixed set of users (tribe) w.r.t. a set of actions. Clearly, tribe leaders are leaders but not vice versa.

Other constraints are added to the framework. The influence emanating from some leaders may be "subsumed" by others. Therefore, in order to rule out such cases Goyal *et al.* introduce the concept of *genuineness*. Finally, similarly to association rules mining, also the constraint of *confidence* is included in the framework.

As observed before, the propagation log might potentially be very large, the algorithmic solution must always try to minimize the number of scans of the propagation log needed. This is fundamental to achieve efficiency. In both the "credit distribution" model [45], and the "leaders and tribes" framework [49], [50], Goyal *et al.* develops algorithms that scan the propagation log only once.

VI. Sparsification of Influence Networks

In this section we review another interesting problem defined over the same input: (1) the social graph, and (2) the log of past propagations.

Given these two pieces of input, assuming the IC propagation model, and assuming to have learned the edge influence probabilities, Mathioudakis *et al.* [47] study the problem of selecting the k most important links in the model, i.e., the set of k links that maximize the likelihood of the observed propagations. Here k might be an input parameter specified by the data analyst, or alternatively k might be set automatically following common model-selection practice. Mathioudakis *et al.* show that the problem is $\mathbf{NP}$-hard to approximate within any multiplicative factor. However, they show that the problem can be decomposed into a number of subproblems equal to the number of the nodes in the network, in particular by looking for a sparsification for the in-degree of each node. Thanks to this observation they obtain a dynamic programming algorithm which delivers the optimal solution. Although exponential, the search space of this algorithm is typically much smaller than the brute force one, but still impracticable for graphs having nodes with a large in-degree.

Therefore Mathioudakis *et al.* devise a greedy algorithm named SPINE (*Sparsification of influence networks*), that achieves efficiency with little loss in quality.

SPINE is structured in two phases. During the first phase it selects a set of arcs D_0 that yields a log-likelihood larger than $-\infty$. This is done by means of a greedy approximation algorithm for the `Hitting Set` $\mathbf{NP}$-hard problem. During the second phase, it greedily seeks a solution of maximum log-likelihood, i.e., at each step the arc that offers the largest increase in log-likelihood is added to the solution set.

The second phase has an approximation guarantee. In fact, while log-likelihood is negative, and not equal to zero for an empty solution, if we consider the gain in log-likelihood w.r.t. the base solution D_0 as our objective function, and we seek a solution of size $k - |D_0|$, then we have a monotone, positive and submodular function g, having $g(\emptyset) = 0$, for which we can apply again the result of Nemhauser *et al.* [40]. Therefore, the solution returned by the SPINE algorithm is guaranteed to be "close" to the optimal among the subnetworks that include the set of arcs D_0.

Sparsification is a fundamental operation that can have countless applications. Its main feature is that by keeping only the most important edges, it essentially highlights the backbone of influence and information propagations in social networks. Sparsifying separately different information topics can help highlighting the different backbone of, e.g., sport or politics. Sparsification can be used for feed ranking [13], i.e., ranking the most interesting feeds for a user. Using the backbone as representative of a group of propagations, can be used for modeling and prototype-based clustering of propagations. Finally, as shown by Mathioudakis *et al.* [47], sparsification can be used as simple data-reduction pre-processing before solving the influence maximization problem. In particular, in their experiments Mathioudakis *et al.* show that by applying SPINE as preprocessor, and keeping only half of the links, Algorithm 1 can achieve essentially the same influence spread σ_m that it would achieve on the whole network, while being an order of magnitude faster.

Another similar problem is tackled by Gomez-Rodriguez *et al.* [52], [53], that assume that the propagations are known, but the network is not. In particular, they assume that connections between nodes cannot be observed, and they use observed traces of activity to infer a sparse, "hidden" network of information diffusion.

Serrano *et al.* [54], as well as Foti *et al.* [55], focus on weighted networks and select edges that represent statistically significant deviations with respect to a null model.

VII. Concluding remarks and open problems

We have provided a brief, partial, and biased survey on the topic of social influence and how it propagates in social networks, mainly focussing on the problem of influence maximization for viral marketing. We have emphasized that while most of the literature has been focussing only on the social graph, it is very important to exploit available traces of past propagations. Finally, we have highlighted the importance of devising clever algorithms to minimize the number of scans of the propagations log.

Although this topic has received a great deal of attention in the last years, many problems remain more or less open.

Learning the strength of the influence exerted from a user of a social network on another user, is a relevant task whose importance goes beyond the mere influence maximization process as depicted in Figure 1. Although some effort has been devoted to investigating this problem (as partially reviewed in Section IV), there is still plenty of room for improving the models and the algorithms for such a learning task.

One important aspect, only touched in [48] is to consider the different levels of user influenceability, as well as the different level of action virality, in the theory of viral marketing and influence propagation. Another extremely important factor is the temporal dimension: nevertheless the role of time in viral marketing is still largely (and surprisingly) unexplored.

We have seen that direct mining methods, as those ones described in Section V, are promising both for what concerns the accuracy and the efficiency in modeling the spread of social influence. In the next years we expect to see more models of this kind.

In a recent paper, Bakshy *et al.* [19] challenge the vision of word-of-mouth propagations that are driven disproportionately by a small number of key influencers. Instead they claim that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, thereby capturing average effects. From this perspective the "leaders and tribes" framework [49], [50] might be an appealing basic brick to build more complex solutions (as it often happens with frequent local patterns which are not very interesting *per se*, but that are very useful to build global models). It would be interesting to see how tribe leaders extracted with the framework of [49], [50] perform when used as seed set in the influence maximization process. Another appealing idea is to use these small tribes as basic units to build larger communities, thus moving towards *community detection based on influence/information propagation*.

The influence maximization problem as defined by Kempe et al. [39] assumes that there is only one player introducing only a product in the market. However, in the real world, is more likely the case where multiple players are competing with comparable products over the same market. Just think about consumers technologies such as videogame consoles (X-Box Vs. Playstation) or reflex digital cameras (Canon Vs. Nikon): as the adoption of these consumers technologies is not free, it is very unlikely that the average consumer will adopt both competing products. Thus is makes sense to formulate the influence maximization problem in terms of mutually exclusive and competitive products. While there are two papers that have tackled this problem independently and concurrently in 2007 [56], [57], their contribution is mostly theoretical and leaves plenty of room for developing more concrete analysis and methods.

One important aspect largely left uncovered in the current literature is the fact that some people are more likely to buy a product than others, e.g., teenagers are more likely to buy videogames than seniors. Similarly, a user which is influential w.r.t. classic rock music, is not very likely to be influential for what concerns techno music too. These considerations highlight the need of, (1) methods that can take benefit of additional information associated to the nodes (the users) of a social network (e.g., demographics, behavioral information), and (2), methods to incorporate topic modeling in the influence analysis. While some preliminary work in this direction exists [58], [18], [59], we believe that the synergy of topic modeling and influence analysis is still in its infancy, and we expect this to become an hot research area in the next years.

Mining influence propagations data for applications such as viral marketing has non-trivial privacy issues. Studying the privacy threats associated to these mining activities and devising methods respectful of the privacy of the social networks users are important problems.

Finally, the main open challenge in our opinion is that the influence maximization problem, as defined by Kempe et al. [39] and as reviewed in this article, is still an ideal problem: how to make it actionable in the real world? Propagation models, e.g., the IC and LT models reviewed in Section II (but many more exist in the literature), make many assumptions: which of these assumptions are more realistic and which are less? Which propagation model does better describe the real-world? We need to develop techniques and benchmarks for comparing different propagation models and the associated influence maximization methods on the basis of ground-truth.

Acknowledgements

I wish to thank Amit Goyal and Laks V. S. Lakshmanan which are my main collaborators in the research on the topic of influence propagation and the co-authors of most of the papers discussed in this article. I would also like to thank Michael Mathioudakis and my colleagues at Yahoo! Research Barcelona: Carlos Castillo, Aris Gionis, and Antti Ukkonen.

I wish to thank Paolo Boldi for helpful discussions and detailed comments on an earlier version of this manuscript.

Finally I would like to thank the chairs and organizers of WI-IAT 2011 (www.wi-iat-2011.org) conference for inviting me to give a keynote, as well as the editors of the IEEE Intelligent Informatics Bulletin for inviting me to summarize the keynote in this article. My research on influence propagation is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (www.cenitsocialmedia.es).

References

[1] J. Coleman, H. Menzel, and E. Katz, *Medical Innovations: A Diffusion Study*. Bobbs Merrill, 1966.

[2] T. Valente, *Network Models of the Diffusion of Innovations*. Hampton Press, 1955.

[3] F. Bass, "A new product growth model for consumer durables," *Management Science*, vol. 15, pp. 215–227, 1969.

[4] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, no. 3, pp. 211–223, 2001.

[5] V. Mahajan, E. Muller, and F. Bass, "New product diffusion models in marketing: A review and directions for research," *Journal of Marketing*, vol. 54, no. 1, pp. 1–26, 1990.

[6] S. Jurvetson, "What exactly is viral marketing?" *Red Herring*, vol. 78, pp. 110–112, 2000.

[7] N. E. Friedkin, *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

[8] J. Wortman, "Viral marketing and the diffusion of trends on social networks," University of Pennsylvania, Tech. Rep. Technical Report MS-CIS-08-19, May 2008.

[9] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[10] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun, "Personalized recommendation driven by information flow," in *Proc. of the 29th ACM SIGIR Int. Conf. on Research and development in information retrieval (SIGIR'06)*, 2006.

[11] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in *Proc. of the 16th Int. Conf. on World Wide Web (WWW'07)*, 2007.

[12] J. J. Samper, P. A. Castillo, L. Araujo, and J. J. M. Guervós, "Nectarss, an rss feed ranking system that implicitly learns user preferences," *CoRR*, vol. abs/cs/0610019, 2006.

[13] D. Ienco, F. Bonchi, and C. Castillo, "The meme ranking problem: Maximizing microblogging virality," in *Proc. of the SIASP workshop at ICDM'10*, 2010.

[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proc. of the 13th Int. Conf. on World Wide Web (WWW'04)*, 2004.

[15] C.-N. Ziegler and G. Lausen, "Propagation models for trust and distrust in social networks," *Information Systems Frontiers*, vol. 7, no. 4-5, pp. 337–358, 2005.

[16] J. Golbeck and J. Hendler, "Inferring binary trust relationships in web-based social networks," *ACM Trans. Internet Technol.*, vol. 6, no. 4, pp. 497–529, 2006.

[17] M. Taherian, M. Amini, and R. Jalili, "Trust inference in web-based social networks using resistive networks," in *Proc. of the 2008 Third Int. Conf. on Internet and Web Applications and Services (ICIW'08)*, 2008.

[18] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proc. of the Third Int. Conf. on Web Search and Web Data Mining (WSDM'10)*, 2010.

[19] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proc. of the Forth Int. Conf. on Web Search and Web Data Mining (WSDM'11)*, 2011.

[20] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. of the 20th Int. Conf. on World Wide Web (WWW'11)*, 2011.

[21] D. M. Romero, B. Meeder, and J. M. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proc. of the 20th Int. Conf. on World Wide Web (WWW'11)*, 2011.

[22] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in *Proc. of the 17th Int. Conf. on World Wide Web (WWW '08)*, 2008.

[23] D. Watts and P. Dodds, "Influential, networks, and public opinion formation," *Journal of Consumer Research*, vol. 34, no. 4, pp. 441–458, 2007.

[24] D. Watts, "Challenging the influentials hypothesis," *WOMMA Measuring Word of Mouth, Volume 3*, pp. 201–211, 2007.

[25] D. Watts and J. Peretti, "Viral marketing for the real world," *Harvard Business Review*, pp. 22–23, May 2007.

[26] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, 2008.

[27] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri, "Feedback effects between similarity and social influence in online communities," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, 2008.

[28] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc. of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 544–21 549, 2009.

[29] T. L. Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," in *Proc. of the 19th Int. Conf. on World Wide Web (WWW'10)*, 2010.

[30] R. Lyons, "The spread of evidence-poor medicine via flawed social-network analysis," *Statistics, Politics, and Policy*, vol. 2, no. 1, 2011.

[31] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *The New England Journal of Medicine*, vol. 357(4), pp. 370–379, 2007.

[32] M. Cha, A. Mislove, and P. K. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proc. of the 18th Int. Conf. on World Wide Web (WWW'09)*, 2009.

[33] J. Leskovec, A. Singh, and J. M. Kleinberg, "Patterns of influence in a recommendation network," in *Proc. of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining,(PAKDD'06)*, 2006.

[34] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *TWEB*, vol. 1, no. 1, 2007.

[35] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, vol. 21, no. 2, pp. 256–276, 2006.

[36] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. of the Seventh ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'01)*, 2001.

[37] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, 2002.

[38] D. M. Chickering and D. Heckerman, "A decision theoretic approach to targeted advertising," in *Proc. of the 16th Conf. in Uncertainty in Artificial Intelligence (UAI'00)*, 2000.

[39] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03)*, 2003.

[40] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions - i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[41] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'10)*, 2010.

[42] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. of the 10th IEEE Int. Conf. on Data Mining (ICDM'10)*, 2010.

[43] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'07)*, 2007.

[44] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09)*, 2009.

[45] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *PVLDB*, vol. 5, no. 1, pp. 73–84, 2011.

[46] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. of the 12th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES'08)*, 2008.

[47] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen, "Sparsification of influence networks," in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, 2011.

[48] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Third ACM Int. Conf. on Web Search and Data Mining (WSDM'10)*, 2010.

[49] ——, "Discovering leaders from community actions," in *Proc. of the 2008 ACM Conf. on Information and Knowledge Management (CIKM 2008)*, 2008.

[50] A. Goyal, B.-W. On, F. Bonchi, and L. V. S. Lakshmanan, "Gurumine: A pattern mining system for discovering leaders and tribes," in *Proc. of the 25th IEEE Int. Conf. on Data Engineering (ICDE'09)*, 2009.

[51] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'93)*, 1993.

[52] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'10)*, 2010.

[53] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML'11)*, 2011.

[54] M. A. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proc. of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[55] N. J. Foti, J. M. Hughes, and D. N. Rockmore, "Nonparametric sparsification of complex multiscale networks," *PLoS ONE*, vol. 6, no. 2, 2011.

[56] S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in *Proc. of the Third Int. Workshop on Internet and Network Economics (WINE'07)*, 2007.

[57] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen, "Maximizing influence in a competitive social network: a follower's perspective," in *Proc. of the 9th Int. Conf. on Electronic Commerce (ICEC'07)*, 2007.

[58] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09)*, 2009.

[59] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. of the 19th ACM Conf. on Information and Knowledge Management (CIKM'10)*, 2010.

Mining a Data Reasoning Model for Personalized Text Classification

Luepol Pipanmaekaporn* and Yuefeng Li†
Computer Science Discipline, Faculty of Science and Technology
Queensland University of Technology, Brisbane, QLD 4001, Australia
Email: luepol.p@gmail.com* and y2.li@qut.edu.au†

***Abstract*—It is a big challenge to acquire correct user profiles for personalized text classification since users may be unsure in providing their interests. Traditional approaches to user profiling adopt machine learning (ML) to automatically discover classification knowledge from explicit user feedback in describing personal interests. However, the accuracy of ML-based methods cannot be significantly improved in many cases due to the term independence assumption and uncertainties associated with them.**

This paper presents a novel relevance feedback approach for personalized text classification. It basically applies data mining to discover knowledge from relevant and non-relevant text and constraints specific knowledge by reasoning rules to eliminate some conflicting information. We also developed a Dempster-Shafer (DS) approach as the means to utilise the specific knowledge to build high-quality data models for classification. The experimental results conducted on Reuters Corpus Volume 1 and TREC topics support that the proposed technique achieves encouraging performance in comparing with the state-of-the-art relevance feedback models.

***Index Terms*—Personalized text Classification, User Profiles, Relevance Feedback, Reasoning Model, and Data Mining**

I. Introduction

As the vast amount of online information available causes information overloading, the demand for personalized approaches for information access increases. One of the key techniques for personalized information access is personalized text classification [1], [4], where a system is able to retrieve or filter contents according to personal interests [9]. As for personalization, a user profile is used to represent user interests and perferences.

It is not uncommon that hand-coding user profiles is impractical since users may be unsure of their interests or not have any technical knowledge to describe their profile. It is hence preferable to directly learn classifiers from examples. A common user profiles acquiring approach is to explore relevance feedback (RF). In particular, a user is given to express his/her opinions by deciding which documents are *relevant* or *non-relevant* to the user. By using the explicit feedback, machine learning (ML) techniques could be adopted to learn a text classifier that represents the user interest [23], [24] or search intent [40]. For example, Rocchio [12], [20] and SVMs [13], [14] are two effective learning algorithms in this literature.

Nevertheless, the performance of ML-based approaches to RF often cannot significantly improve. This is since the nature of ML techniques that require a large training set to achieve good performance whereas in fact the number of feedback documents given by a user is small. Furthermore, ML-based approaches typically deal with training documents with the term independence assumption and ignore any syntactic and semantic information of correlations between terms. As a result, they may miss some useful terms that are added into user profiling models [3], [6], [21].

Data mining (DM) based approaches to relevance feedback have recently given great interests [34], [35], [39]. These approaches basically discover frequent patterns that capture frequent terms and their relationships in text and consequently utilise the discovered patterns for constructing relevance models. In [35], the authors adopted data mining to mine relevant documents in order to discover a set of sequential patterns. A document evaluation function is formed by those patterns to score new documents. A *deploying* method was proposed [34] to solve the problem of low-frequency occurrence of patterns in text. Instead of patterns, a weighted vector of terms is generated by discovered patterns and used for building the relevance model. Some deploying-based approaches (e.g. IPE [39]) attempt to improve the quality of relevance model by using negative feedback. Although experimental results conducted on RCV1 data collection illustrate the usefulness of frequent patterns for personalized text search, we believe that the existing approaches may not be able to obtain high-quality relevance feedback models. Firstly, these approaches focus on building relevance models that use patterns, but ignore the attempt to select a small set of high-quality patterns. Furthermore, it is still not clear how to effectively deal with the result of pattern mining to improve the effectiveness of relevance feedback models.

Motivated by these issues, this paper presents a novel relevance feedback approach for discovering user profiles from text using data mining and reasoning techniques. In specific terms, it discovers features (frequent patterns) from relevant and non-relevant text and constraints specific ones by reasoning rules to eliminate some conflicting information. To construct the user profile model, we developed a Dempster-Shafer (DS) approach that allows to establish the connection between patterns and terms. It also allows to incorporate the uncertain nature of text features (i.e., terms and patterns) for modelling user's interests. The experimental results conducted on RCV1 and TREC text collections [22] support that the data reasoning approach achieves encouraging performance as compared to the state-of-the-art techniques.

TABLE I
A SET OF PARAGRAPHS

$Paragraph$	$Terms$
dp_1	$t_1\ t_2$
dp_2	$t_3\ t_4\ t_6$
dp_3	$t_3\ t_4\ t_5\ t_6$
dp_4	$t_3\ t_4\ t_5\ t_6$
dp_5	$t_1\ t_2\ t_6\ t_7$
dp_6	$t_1\ t_2\ t_6\ t_7$

In summary, our contributions include

- We propose a novel relevance feedback approach for personalized text categorization.
- We analysis text patterns by observing their semantic relationships and devising reasoning rules that investigate specific patterns to describe user interests.
- We propose a novel method for constructing user profiles using frequent patterns in text to improve performance of text categorization.

The rest of the paper is organized as follows: Section 2 gives some basic definitions of frequent patterns in text. In section 3, we provide a data mining framework for discovering features in relevant and non-relevant text. We also describe a novel feature selection method based on the investigation of reasoning rules. Section 4 presents how Dempster-Shafer approach facilitate the utilisation of the discovered patterns for constructing the user profile model. Extensive experimental results are presented in Section 5 and related work is discussed in Section 6, following by conclusions in Section 7.

II. BACKGROUND

Let D be a training set of documents, including a set of positive (relevant) documents, D^+, and a set of negative (irrelevant) ones, D^-. When splitting a document into paragraphs, a document d can also be represented by a set of paragraphs $PS(d)$.

A. Frequent and Closed Patterns

Let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of terms which are extracted from D^+. Given X be a set of terms (called a *termset*) in document d, $coverset(X)$ denotes the covering set of X for d, which includes all paragraphs $dp \in PS(d)$ where $X \subseteq dp$, i.e., $coverset(X) = \{dp|dp \in PS(d), X \subseteq dp\}$. The *absolute* support of X is the number of occurrences of X in $PS(d)$: $sup_a(X) = |coverset(X)|$. The*relative* support of X is the fraction of the paragraphs that contain the pattern: $sup_r(X) = \frac{|coverset(X)|}{|PS(d)|}$. A termset X called *frequent pattern* if its sup_a (or sup_r) $\geq min\_sup$, a minimum support.

Table I lists six paragraphs for a given document d, where $PS(d) = \{dp_1, dp_2, \ldots, dp_6\}$, and duplicate terms are removed. Assume $min\_sup = 3$, ten frequent patterns would be extracted as shown in Table II.

Given a set of paragraphs $Y \subseteq PS(d)$, we can define its *termset*, which satisfies

$$termset(Y) = \{t|\forall dp \in Y \Rightarrow t \in dp\}$$

By defining the closure of X as:

$$Cls(X) = termset(coverset(X))$$

TABLE II
SEQUENTIAL PATTERNS AND COVERING SETS

$Frequent\ Pattern$	$Covering\ Set$
$\{\mathbf{t_3}, \mathbf{t_4}, \mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{\mathbf{t_1}, \mathbf{t_2}\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{\mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_5\}$

a pattern (or termset) X is *closed* if and only if $X = Cls(X)$. Let X be a closed pattern. We have

$$sup_a(X_1) < sup_a(X) \qquad (1)$$

for all patterns $X_1 \supset X$.

B. Closed Sequential Patterns

A *sequential pattern* $X =< t_1, \ldots, t_r >$ ($t_i \in T$) is an ordered list of terms, where its $sup_r \geq min\_sup$. A sequence $s_1 =< x_1, \ldots, x_i >$ is a *sub-sequence* of another sequence $s_2 =< y_1, \ldots, y_j >$, denoted by $s_1 \sqsubseteq s_2$, iff $\exists j_1, \ldots, j_i$ such that $1 \leq j_1 < j_2 \ldots < j_i \leq j$ and $x_1 = y_{j_1}, x_2 = y_{j_2}, \ldots, x_i = y_{j_i}$. Given $s_1 \sqsubseteq s_2$, we usually say s_1 is a *sub-pattern* of s_2, and s_2 is a *super-pattern* of s_1. To simplify the explanation, we refer to sequential patterns as patterns.

As the same as those defined of normal patterns, we define the *absolute support* and *relative support* for a pattern (an ordered termset) X in d. We also denote the covering set of X as $coverset(X)$, which includes all paragraphs $ps \in PS(d)$ such that $X \sqsubseteq ps$, i.e., $coverset(X) = \{ps|ps \in PS(d), X \sqsubseteq ps\}$. X is then called a $frequentpattern$ if $sup_r(X) \geq min\_sup$. By using Eq. (1), a frequent sequential pattern X is *cloesd* if not $\exists$ any super-pattern X_1 of X such that $sup_a(X_1) = sup_a(X)$.

To improve the efficiency of finding all closed sequential patterns from training documents, an algorithm, $SPMining(D^+, min\_sup)$, was introduced by [35]. The *SPMining* algorithm used well-known *Apriori* property to narrow down the searching space.

C. Dempster-Shafer theory

Dempster-Shafer (hereafter DS) [37] is a statistically based technique for combining evidence. It can be considered a generalization of Bayesian theory as it allows assignment of probability to uncertain events, offering a way to represent ignorance or uncertainty. A beneficial characteristic of DS is the ability to use partial knowledge over propositions and represent uncertainty as part of a modelling process. Recently, there are an increasing number of developments and applications using DS approach. In particular, generalized evidence theory [10], [18], Data Fusion [32], Machine Learning [7], [8], association rules mining [17], Information Retrieval [26], [30], Web mining models [15], [36].

In general terms, DS deals with a finite set of exclusive and exhaustive propositions, called the *frame of discernment* (denoted by Ω). All the subsets of Ω belong to the power set of Ω, denoted by 2^{Ω}. A strength of subset of elements in Ω is given by the definition of a mass function $m : 2^{\Omega} \rightarrow [0, 1]$, which provides a measure of uncertainty, applied over all the subsets of elements in the frame of discernment. The mass function also satisfies the following properties:

(1) $m(\emptyset) = 0$ and

(2) $\sum\limits_{A \in 2^{\Omega}} m(A) = 1$

DS provides a rule, known as the Dempster's rule of combination [25], for combining evidence, possibly originating from different sources of data (e.g. Sensors). The combination yields a probability mass assigned to a subset of Ω, given a subset of propositions A, characterized by a mass distribution m_1 and subset of propositions B, characterized by a mass distribution m_2. The normalized version of the combination rule is the following:

$$m_1 \oplus m_2(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)} \tag{2}$$

for all $A \in 2^{\Omega}$, where $m_1 \oplus m_2(A)$ denotes the combine evidence.

In the DS, probability masses applied over all the subsets of elements in the frame of discernment can be used to infer the mass for the single elements as the means to make decisions. The masses are represented by probability functions called *Pignistic* probabilities [27]. The pignistic probability is defined as:

$$BetP(A) = \sum_{B \subseteq \Omega} \frac{|A \cap B|}{|B|} \frac{m(B)}{(1 - m(\emptyset))} \tag{3}$$

for all subsets $A \subseteq \Omega$. A shortcoming of DS is related to the use of masses instead of probability measures and high involvements for users to explicitly provide values for the mass functions.

A shortcoming of DS is related to the use of masses instead of probability measures and difficulties in coming up with these values for the mass functions [37].

III. Relevant Feature Discovery

Vector Space Model (VSM) is the popular choice for representing information in text documents since it is efficient and effective for text processing. However, it fails to capture semantic information which is often represented by relations between terms (i.e., syntactic or semantic phrases). Finding phrases in text is related to mining frequent subsequences in sequence collections. We hence apply data mining to discover useful features available in relevant text. Mining sequential patterns offers to generate both low-level features (terms) and high-level ones (phrases) in sentences or paragraphs w.r.t. frequency. They enjoy statistical properties since they are frequent. Furthermore, many noisy patterns could be removed w.r.t. the minimum support constraint. We adopt $SPMining$ algorithm [35] (also used in [34] [39]) to discover frequent subsequences (hereafter patterns) in paragraphs of positive documents D^{+}. For all positive documents $d_i \in D^{+}$, the *SPMining* algorithm finds a set of patterns based on a given min_sup to obtain the following vector:

$$\overrightarrow{d_i} = \langle (p_{i_1}, f_{i_1}), (p_{i_2}, f_{i_2}), \ldots, (p_{i_m}, f_{i_m}) \rangle \tag{4}$$

where p_j in pair (p_j, f_j) denotes a pattern and f_j is its frequency in d_i. The result of this algorithm is a set of document vectors, which can be expressed as follows.

$$\eta = \left\{ \overrightarrow{d_1}, \overrightarrow{d_2}, \ldots, \overrightarrow{d_n} \right\}$$

where $n = |D^{+}|$.

A. A weighted combination operator

For each vector $\overrightarrow{d_i} \in \eta$, the frequency of pattern can imply the pattern's significance in the context of document. As training documents may contain a pattern more than once, it is important to determine which patterns are significant in aspects of information use. However, existing data mining algorithms usually exclude the local support information by only considering their binary presence and absence in training documents. As a result, they lose in the local pattern's significance that may provide some insights. For example, considering two patterns p and q that occur 20 times and 2 times in the same document with equal importance can be incorrect.

To achieve this, we apply the idea of data fusion to effectively combine multiple sets of patterns in different documents into a single one. In information retrieval, data fusion has been used to combine results from different retrieval models, different document representations, different query representations and so on, to improve effectiveness [30], [33].

We first define a score function ρ_i that assigns a score to a pattern p_j based on its frequency in a document d_i as the following equation:

$$\rho_i(p_j) = \begin{cases} \frac{f_j}{\sum_{p_k \in \overrightarrow{d_i}} f_k} & ; p_j \in \overrightarrow{d_i} \\ 0 \quad ; otherwise \end{cases} \tag{5}$$

where f_j denotes the absolute support value (Sup_a) of pattern p_j in document d_i. Given two score functions ρ_a and ρ_b belonging to document d_a and d_b respectively, we define a weighted linear combination operator $\oplus$ to compose the two score functions into the combined score for pattern p_j. This operator can be found as the following equation:

$$\rho_a \oplus \rho_b(p_j) = \frac{1}{K} \times \begin{cases} w_a \times \rho_a(p_j) + w_b \times \rho_b(p_j) & ; p_j \in \overrightarrow{d_a} \cap \overrightarrow{d_b} \\ w_a \times \rho_a(p_j) \quad ; p_j \notin \overrightarrow{d_b} \\ w_b \times \rho_b(p_j) \quad ; p_j \notin \overrightarrow{d_a} \\ 0 \quad ; p_j \notin \overrightarrow{d_a} \cup \overrightarrow{d_b} \end{cases} \tag{6}$$

where w_a and w_b be a user-defined weight associated with document d_a and d_b respectively and $K = w_a + w_b$, which is a weight normalisation. The weights reflect the importance of feedback documents, which can be the document's length or the degree of perceived relevance given by a user or IR system. If all the documents are equally weighted, then the combined score of pattern is fairly averaged.

Let SP^+ be a set of sequential patterns collected from all the relevant training documents in a collection D^+, i.e., $SP^+ = Sp_1 \cup Sp_2 \cup \cdots \cup SP_{|D^+|} = \bigcup_{i=1}^{|D^+|} Sp_i$. For each pattern $p_j \in SP^+$, the score assigned to the pattern p_j can calculated by combining all the score functions of documents in a document collection as the following equation:

$$\rho_c(p_j) = \bigoplus_{i=1}^{|D^+|} \rho_i = \frac{1}{K} \sum_{i=1..|D^+|} w_i \times \rho_i(p_j) \qquad (7)$$

where $\rho_c(p_j)$ returns the combined support given to the pattern p_j, w_i is a weight associated with document d_i and $K = w_1 + w_2 + \cdots + w_{|D^+|}$. Since we only know which documents are positive or negative, but not which one is more important, in this paper all training documents are equally treated (i.e., 1).

B. From relevant features to specific features

Although patterns provide highly detailed descriptors for document representation, their large number of generated patterns may hinder their effective use. This is since many of these patterns are redundant and conflict. Adding such patterns can harm the classification accuracy due to the overfitting effect; however, it is very difficult to identify which patterns are noisy since they depend on users' perspectives for their information needs [39].

To this end, we propose a novel method to detect and eliminate patterns that are conflict. The idea is to check which patterns have been used in the context of the non-relevant data. We first define two kinds of errors: *total conflict* error and *partial conflict* error.

Definition 1 (total conflcit). *Given a pattern $p \in R$, p is called total conflict with a category $\overline{R}$ if $\exists q \in \overline{R}$ and $termset(p) \subseteq termset(q)$.*

Definition 2 (partial conflcit). *Given a pattern $p \in R$, p is called partial conflict with a category $\overline{R}$ if $\exists q \in \overline{R}$ and $termset(p) \cap termset(q) \neq \emptyset$.*

To apply this idea, we discover patterns from negative documents in D^- and consequently fuse them into a single collection, defined as SP^-, as we did in positive documents. Based on the above definitions, we identify all patterns in response to the following rules:

$$\begin{aligned} &S^+ = \{p | p \in SP^+, \forall q \in SP^- \Rightarrow p \nsubseteq q\} \\ &S^- = \{q | q \in SP^-, \exists p \in SP^+ \Rightarrow q \cap p \neq \emptyset\} \\ &N = (SP^+ \cup SP^-) - S^+ - S^- \end{aligned}$$

where $S^+ \cap S^- \cap N = \emptyset$. SP^+ and SP^- are two categories of patterns in the relevant and non-relevant data respectively. For the relevant category, non-conflict patterns contain termsets that are specific to user's interests or a user because they never overlap with any patterns from the non-relevant category while partial conflict ones are termsets that share a part with some of those patterns. All these patterns are classified into S^+. On the other hand, total conflict patterns in the relevant category are classifed into N since they contain termsets that may frequently occur in the context of non-relevant data.

A collection S^- consists of all conflict patterns in the non-relevant category. Such patterns are useful to identify noisy terms in the relevant data. Also, non-conflict patterns in the non-relevant category is classified into N because they are irrelevant data.

Once patterns were classified, we store all patterns in S^+ and S^- and remove patterns contained in the collection N.

IV. User Profile Construction

In this section, we describe our approach to construct the user profile model by using the specific knowlege.

A. Mapping patterns to belief functions

The initial user profile is first built based on two category of patterns S^+ and S^-. Let Ω consists of n terms extracted from all patterns in the two pattern collections, i.e., $\Omega = \{t_1, t_2, \ldots, t_n\}$. We define a set-valued mapping $\psi :: S^+ \cup S^- \to 2^\Omega$ to associate the relationship between patterns and the term space in Ω to generate mass functions.

Based on this mapping, we define a mass function $m^+ : 2^\Omega \to [0, 1]$ on Ω, the set of terms, called *positive mass* function, which satisfies

$$m^+(A) = \left\{ \begin{array}{cl} 0 \quad if \quad A = \emptyset; & \\ \frac{\rho_c(\{p|p \in S^+, \Gamma(p)=A\})}{\sum_{B \subseteq \Omega} \rho_c(\{q|q \in S^+, \Gamma(q)=B\})} & , otherwise \end{array} \right. \qquad (8)$$

for all $A \subseteq \Omega$, where $\rho_c(p)$ returns the combined support of pattern p obtained by Eq.(5) and B is a subset on Ω space.

As we did in the positive data, patterns from non-relevant data (i.e., S^-) can be used to generate mass functions, defined as *negative mass* functions ($m^-(A)$). A positive mass function $m^+(A)$ represents the strength that supports set A, a set of terms, whereas a negative mass function $m^-(A)$ means the contrary.

Figure 1 illustrates an example of mapping knowledge (patterns) to mass functions on Ω space.

A main advantage of representing the discovered knowledge with belief functions is that uncertainties associated with text features (i.e., frequent terms and patterns) can be represented.

B. Weight assignment by belief functions

In order to reason with the derived mass functions, we present the new idea to assign weights of terms in the profile vector. The main advantage of the weight assignment method is that it takes uncertainties represented by mass functions in estimating term weights.

For each term $t_i \in \Omega$, we first transfer positive mass functions into a *pignistic probability* [28] as the following functions:

$$Pr_{m^+}(t_i) = \sum\nolimits_{\emptyset \neq A \subseteq \Omega, t_i \in A} \frac{m^+(A)}{|A|} \qquad (9)$$

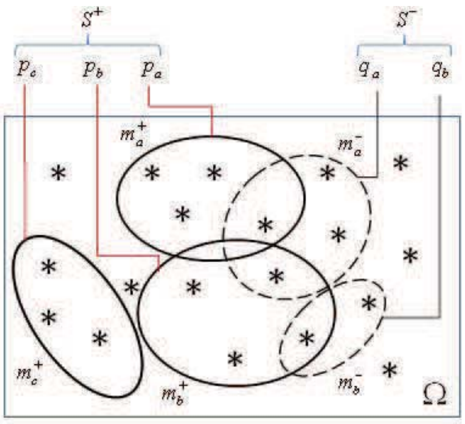


Fig. 1. An example of mass functions generated by discovered patterns

where A denotes a subset of elements in Ω and $|A|$ is the number of elements in A. The pignistic value represents the expected probability assigned to single elements in the frame of discernment for betting [28]. In our case, we use the resulting probability as the means to score terms in the profile vector corresponding to their distribution in the term dependency data (i.e, positive and negative data). The high value assigned to a term represents the high importance of the term in the underlying data. The pignisitic probability assigned to a term t_i with negative mass functions can be estimated by

$$Pr_{m^-}(t_i) = \sum_{\emptyset \neq A \subseteq \Omega, t_i \in A} m^-(A)/|A|.$$

Finally, the two probability functions are combined to estimate the weight of each term t_i in the user profile Ω as the following equation:

$$w(t_i) = \frac{Pr_{m^+}(t_i) \times (1 - Pr_{m^-}(t_i))}{1 - \min\{Pr_{m^+}(t_i), Pr_{m^-}(t_i)\}} \qquad (10)$$

The term's weight measures the term's importance in respect to the user's interests. When the pignistic value of a term given by positive mass functions ($Pr_{m^+}(t_i)$) is high, the term tends to be a good identifier for identify relevant documents. As a result, the term weight value tends to be high. Conversely, the pignistic probability with negative mass functions ($Pr_{m^-}(t_i)$) is supposed to be negatively correlated with the user's topic of interest. When this value is high indicating that the term tend to be used in describing other topics, the important weight given to the term is reduced as a consequence.

A document evaluation function is built for the use of user profile in document filtering. Given a new document d, the relevance score given to the document d can be calculated as the following function:

$$R(d) = \sum_{t \in d} \frac{tf(t)}{\sum_{t_j \in d} tf(t_j)} \times support(t) \qquad (11)$$

where $support(t) = w(t)$ if $t \in \Omega$; otherwise $support(t) = 0$ and $tf(t)$ denotes the term frequency of term t in document d.

It is easy to apply a threshold strategy to the document evaluation function for making a binary decision, aiming to predict the class labels of document d into relevant and non-relevant to a user. Given a threshold value ζ, if $r(d) \geq \zeta$ then the document d is *relevant*; otherwise it is *non-relevant*. The best value of ζ can emperically estimated.

V. Experimental Evaluation

In this section, we evaluate the proposed approach. We conduct experiments on RCV1 data collection and TREC topics. We also discuss the testing environment including the data collection, baseline models, and evaluation methods. The data reasoning model (afterhere *DRM*) is a supervised approach that needs a training set including both positive (relevant) documents and negative (non-relevant) documents from an individual user.

A. RCV1 data collection and TREC topics

Reuter Corpus Volume 1 (RCV1) is used to test the effectiveness of the proposed model. RCV1 corpus consists of all and only English language stories proposed by Reuter's journalists between August 20,1996, and August 19,1997, a total of 806,791 documents that cover very large topics and information [22]. For each topic, some documents in RCV1 data collection are divided into a training set and a testing set. TREC(2002) has developed and provided 50 assessor topics for the filtering track, aiming at building a robust filtering system [29]. The relevance judgements of documents in the assessor topics have been made by human assessors of the National Institute of Standards and Technology (NIST),i.e., assessor topics. According to [29], the justification of enough using the 50 assessor topics for evaluating robust IF systems was given. In this study, we use only the 50 assessor topics for performance evaluation in the proposed model.

All documents in RCV1 are marked in XML. To avoid bias in the experiments, all the meta-data information in the collection have been ignored. The documents are treated as plain text documents by preprocessing the documents. The tasks of removing stop-words according to a given stop-words list and stemming terms by applying the Porter Stemming algorithm are applied.

B. Baseline Models and Settings

We group baseline models into two main categories. The first category includes a number of data mining (DM) based methods for IF (i.e., PTM [35], PDS [34] and IPE [39] while the second category includes two effective machine learning models in text categorization and filtering (i.e. Rocchio [12] and SVM [13]). DM-based models were discussed in the section Related work.

1) DM-based models:: Both PTM and PDS models use only positive features (i.e, patterns for the case of PTM and terms for the case PDS) from training relevant documents to generate user profiling models while IPE uses both positive and negative features. For data mining models, the minimum support threshold (min_sup) is an important parameter and is sensitive for a specified data set. We set $min\_sup = 0.2$ (20% of number of paragraphs in a document) for all baseline models in the category (also DRM) since this value was recommended best for this data collection [34], [35], [39].

2) Machine Learning based models:: The Rocchio algorithm has been widely adopted in text categorization and filtering [12], [24]. The Rocchio builds a Centroid for representing user profiles. The centroid $\vec{c}$ of a topic can be generated as follows:

$$\alpha\frac{1}{|D^{+}|}\sum_{\vec{d}\in D^{+}}\frac{\vec{d}}{||\vec{d}\,||}-\beta\frac{1}{|D^{-}|}\sum_{\vec{d}\in D^{-}}\frac{\vec{d}}{||\vec{d}\,||} \tag{12}$$

where $||\vec{d}\,||$ be normalized vector for document d. α and β be a control parameter for the effect of relevant and nonrelevant data respectively. According to [5], [12], there are two recommendations for setting the two parameters: $\alpha = 16$ and $\beta = 4$; and $\alpha = \beta = 1.0$. We have tested both accommodations on assessor topics and found the latter recommendation was the best one. Therefore, we let $\alpha = \beta = 1.0$.

SVM is a state-of-the-art classifier [13]. In our experiments, we used the linear kernel since it has been proved to be as powerful as other kernels when tested on the Reuters-21578 data colleciton for text classification [24]. We hence used the following decision function in SVM:

$$h(x) = sign(w.x+b) = \begin{cases} +1 & if(w.x+b) > 0 \\ -1 & otherwise \end{cases}$$

where x is the input object; $b \in R$ is a threshold and $w = \sum_{i=1}^{l} y_i\alpha_i x_i$ for the given training data:$(x_i, y_i), \ldots, (x_l, y_l)$, where $x_i \in R^n$ and $y_i = +1(-1)$, if document x_i is labeled positive (negative). $\alpha_i \in R$ is the weight of the sample x_i and satisfies the constraint:

$$\forall_i: \quad \alpha_i \geq 0 \quad and \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{13}$$

The SVM here is used to rank documents rather than to make a binary decision, and it only uses terms based features extracted from training documents. For this purpose, threshold b can be ignored. For the documents in a training set, we know only what are positive (negative), but not which one is important. We assign the same α_i value (i.e., 1) to each positive document first, and then determine the same α_i (i.e., α') value to each negative document based on the Eq. (11). Therefore, a testing documents d is scored by the function $r(d) = w.d$ where . means *inner products*; d is the term vector of the testing document; and

$$w = \left(\sum_{d_i\in D^{+}} d_i\right) + \left(\sum_{d_j\in D^{-}} d_j\alpha'\right) \tag{14}$$

For each assessor topic, we choose 150 terms in the positive documents, based on *tf*idf* values for all ML-based methods.

C. Results

The effectiveness is determined by five different measures commonly used in Information Retrieval (IR): The average precision of the top 20 documents ($top-20$), F_1 measure, Mean Average Precision (MAP), the break-even point (b/p), and Interpolated Average Precision (IAP) on $11-$points. Precision (p), Recall (r), and F_1 are calculated by the following functions:

$$p = \frac{TP}{TP+FP}, r = \frac{TP}{TP+FN}, F_1 = \frac{2*p*r}{p+r}$$

where TP is the number of documents the system correctly identifies as positives; FP is the number of documents the system falsely identifies as positives; FN is the number of relevant documents the system fails to identify. The larger a $top-20$, MAP, b/p, F_1 measure score is, the better the system performance. $11-$points measure is also used to compare the performance of different systems by averaging precisions at 11 standard recall values (i.e., recall = 0.0, 0.1,...,1.0).

DRM is firstly compared with all data mining based models. We also compare DRM with the state-of-the-art machine learning based models underpinned by Rocchio and SVM for each measuring variable over all the 50 assessing topics.

TABLE III
COMPARISON RESULTS OF DRM WITH ALL DM-BASED METHODS ON ALL ASSESSOR TOPICS

Model	*top-20*	*MAP*	*b/p*	$F_{\beta=1}$
DRM	**0.549**	**0.484**	**0.470**	**0.466**
PTM (IPE) [39]	0.493	0.441	0.429	0.440
PTM (PDS) [34]	0.496	0.444	0.430	0.439
PTM (Closed Seq. ptns) [35]	0.406	0.364	0.353	0.390
%chg	+11.35	+9.75	+9.55	+5.90

TABLE IV
COMPARISON RESULTS OF DRM WITH ALL ML-BASED METHODS ON ALL ASSESSOR TOPICS

Model	*top-20*	*MAP*	*b/p*	$F_{\beta=1}$
DRM	**0.549**	**0.484**	**0.470**	**0.466**
Rocchio [12]	0.474	0.431	0.420	0.430
SVM	0.453	0.408	0.421	0.409
%chg	+15.82%	+12.29%	+11.90%	+8.37%

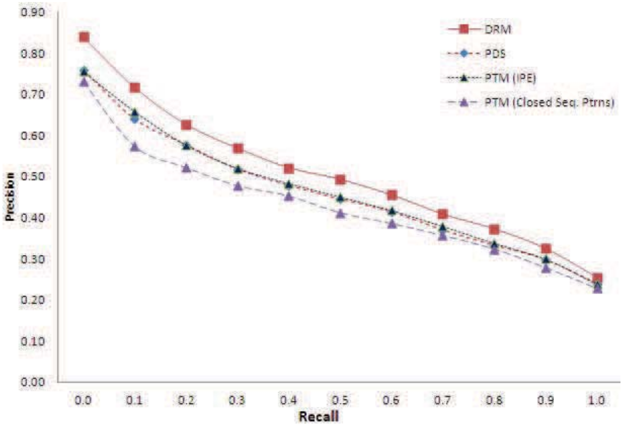


Fig. 2. Comparison results of DRM with all DM-based methods in IAP $11-$points

1) DRM vs data mining-based models: The results of overall comparisons between DRM and all DM based models have shown in Table III. The most important findings revealed in this table are that both PDS and IPE models outperforms

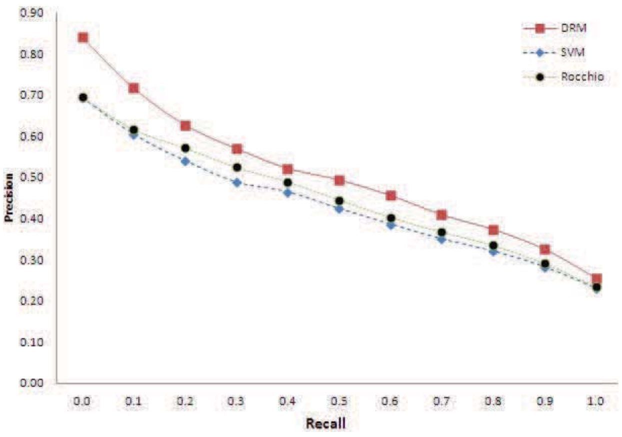


Fig. 3. Comparison results of DRM with all ML-based methods in IAP 11−points

PTM model over all the standard measures while the slight increase in IPE as compared to PDS. The results support the effective use of patterns in text for user profling.

We also compare DRM with IPE. As seen in Table III, DRM significantly incrases for all the evaluation measures with $+9.14\%$ (max $+11.35\%$ on $top - 20$ and min $+5.90\%$ on F_1) in percentage change on average over the standard measures. The encouraging improvments of DRM is also consistent and significant on $11-$points as shown in Figure 1. The results illustrate the highlights of the DS approach to reduce uncertainties involved in estimating term weights.

2) DRM vs machine learning-based models: As shown in Table IV, both Rocchio and SVM models that are based on keyword-based models perform over PTM model, excepting for PDS and IPE. This illustrates keywords remain the very effective concept for text retrieval. However, the results compared between the ML-based models and IPE (also PDS) support that patterns are much effective to select useful terms.

In comparisons with Rocchio and SVM, DRM performs better than Rocchio with $+12.09\%$ increasing in average (max $+15.82\%$ on $top-20$ and min $+8.37\%$ on F_1). The excellent performance of DRM is also obtained as compared to SVM.

VI. Related Work

The frequent pattern-based text classification has been explored by many studies. Earlier approaches are related to associative classification, such as ARC-BC [2], SPAM [11], and HARMONY [31], which mines predictive association rules from a training collection of documents and builds a rule-based text classifier. The results in [2], [38] showed that ARC-BC can perform well on ten most populated Reuters categories as compared to well-known text classifiers, including C4.5, Rocchio, Naive Bayes, excepting for SVMs. In [11], SPAM built by sequential patterns instead of frequent ones showed that it outperformed SVMS in some text collections. HARMOMY [31] focuses on selecting the highest-confidence rules for each training instance to build the text classifier. The objective of our work is different because we are mainly interested in using frequent patterns to build a global model for text classification.

Recently, the focus was more on using frequent patterns to construct new features to improve the quality of text classifier. In [35], a centroid-based text classifier, called PTM, is built by weighted sequential patterns discovered from a relevant text collection. Instead of a full set of features, the closed set is applied to reduce the number of generated patterns. In [19], the authors focused to select top-k discriminative patterns for each training instance from a set of size-1 and size-2 frequent patterns to improve the quality of text classifier. The experimental results showed in [19] highlight the importance of selecting a subset of high-quality patterns.

Nevertheless, the usefulness of frequent patterns is limited by the fact that many mined patterns are never used, especially long patterns. A deploying method for the effective use of patterns in text was proposed in [34], called PDS. It builds a weighted vector of terms from a set of sequential patterns to score new documents corresponding to a relevant category. The result in [16], [34] showed that PDS can largely improve the performance as compared with state-of-the-art text classifiers. Some deploying-based approaches, i.e., IPE [39], focused to improve the classification accuracy by incorporating negative feedback to reduce the effect of noisy terms in relevant documents.

Our work is different from the proposed approaches in the following aspects: (1) we focus on selecting specific patterns from sets of sequential patterns in relevant and non-relevant text collections that describe a target category, where such patterns are investigated by specifying reasoning rules; and (2) we provide a new solution to deal with the set of specific features for text classification. It adopts Dempster-Shafer theory that allows to build the relationship between patterns and terms in estimating weights of terms used in a text classifier.

VII. Conclusions

The paper presents a data reasoning approach for Web user profiling. We have presented a unified model for representing and reasoning about user preference data to construct a correct user profile. We discover patterns from the user data and show how to utilise the patterns for profile construction. We also developed a Dempster-Shafer approach as the means to reduce uncertainties included in text features. Many experiments are conducted on TREC standard text collections and compare the proposed approach with the-state-of-the-art information filtering models. The experiment results illustrate that our proposed approach can improve the system performance.

In the future direction, we are working on reasoning with representations of co-occurrence relations patterns to improve the performance of the data reasoning model.

References

[1] I. Antonellis, C. Bouras, and V. Poulopoulos. Personalized news categorization through scalable text classification. *Frontiers of WWW Research and Development-APWeb 2006*, pages 391–401, 2006.

[2] Maria-Luiza Antonie and Osmar R. Zaïane. Text document categorization by term association. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 19–, Washington, DC, USA, 2002. IEEE Computer Society.

[3] Jing Bai, Jian-Yun Nie, Guihong Cao, and Hugues Bouchard. Using query contexts in information retrieval. In *Proceedings of the 30th international ACM SIGIR Conf.*, pages 15–22, 2007.

[4] A. Baruzzo, A. Dattolo, N. Pudota, and C. Tasso. A general framework for personalized text classification and annotation. *Adaptation and Personalization for Web 2.0*, page 31, 2009.

[5] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *ACM SIGIR 17th International Conf.*, pages 292–300, 1994.

[6] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st international ACM SIGIR Conf.*, pages 243–250. ACM, 2008.

[7] T. Denoeux. A k-nearest neighbor classification rule based on dempster-shafer theory. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 737–760, 2008.

[8] Z. Elouedi, K. Mellouli, and P. Smets. Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, 28(2-3):91–124, 2001.

[9] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The Adaptive Web*, pages 54–89, 2007.

[10] J.W. Guan and D.A. Bell. Evidence theory and its applications, vol. 2. 1992.

[11] S. Jaillet, A. Laurent, and M. Teisseire. Sequential patterns for text categorization. *Intelligent Data Analysis*, 10(3):199–214, 2006.

[12] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of the 4th International Conf. on Machine Learning*, ICML '97, pages 143–151, 1997.

[13] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.

[14] T. Joachims. Transductive inference for text classification using support vector machines. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 200–209. MORGAN KAUFMANN PUBLISHERS, INC., 1999.

[15] Y. Li and N. Zhong. Web mining model and its applications for information gathering. *Knowledge-Based Systems*, 17(5-6):207–217, 2004.

[16] Yuefeng Li, A. Algarni, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceeding of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 753–762, 2010.

[17] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, pages 47–55, 2000.

[18] Dayou Liu and Yuefeng Li. The interpretation of generalized evidence theory. *Chinese Journal of Computers*, 20(2):158–164, 1997.

[19] H.H. Malik and J.R. Kender. Classifying High-Dimensional Text and Web Data using Very Short Patterns. In *8th IEEE ICDM International Conference on Data Mining*, pages 923–928, 2008.

[20] Y.Q. Miao and M. Kamel. Pairwise optimized rocchio algorithm for text categorization. *Pattern Recognition Letters*, 2010.

[21] Nikolaos Nanas and Manolis Vavalis. A ”bag” or a ”window” of words for information filtering? In *Proceedings of the 5th Hellenic Conf. on Artificial Intelligence*, pages 182–193. Springer-Verlag, 2008.

[22] T.G. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1-from yesterdays news to tomorrows language resources. In *3th International Conf. on Language Resources and Evaluation*, pages 29–31, 2002.

[23] S. Schiaffino and A. Amandi. Intelligent user profiling. In *Artificial intelligence*, pages 193–216, 2009.

[24] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, March 2002.

[25] K. Sentz and S. Ferson. Combination of evidence in dempster-shafer theory. Technical report, Citeseer, 2002.

[26] L. Shi, J.Y. Nie, and G. Cao. Relating dependent indexes using dempster-shafer theory. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 429–438. ACM, 2008.

[27] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In *Uncertainty in artificial intelligence*, volume 5, pages 29–39. Elsevier, 1990.

[28] P. Smets. Decision making in a context where uncertainty is represented by belief functions. *Belief functions in business decisions*, 17:61, 2002.

[29] I. Soboroff and S. Robertson. Building a filtering test collection for trec 2002. In *Proceedings of the 26th international ACM SIGIR conference*, page 250. ACM, 2003.

[30] T. Tsikrika and M. Lalmas. Combining evidence for relevance criteria: a framework and experiments in web retrieval. *Advances in Information Retrieval*, pages 481–493, 2007.

[31] J. Wang and G. Karypis. On mining instance-centric classification rules. *IEEE transactions on knowledge and data engineering*, pages 1497–1511, 2006.

[32] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang. Sensor fusion using dempster-shafer theory. In *IEEE Instrumentation and Measurement Technology Conference Proceedings*, volume 1, pages 7–12. Citeseer, 2002.

[33] S. Wu and S. McClean. Performance prediction of data fusion for information retrieval. *Information processing & management*, 42(4):899–915, 2006.

[34] S.T. Wu, Y. Li, and Y. Xu. Deploying approaches for pattern refinement in text mining. In *6th IEEE ICDM International Conf. on Data Mining*, pages 1157–1161, 2006.

[35] S.T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen. Automatic pattern-taxonomy extraction for web mining. In *3th IEEE/WIC/ACM WI International Conf. on Web Intelligence*, pages 242–248, 2004.

[36] Y. Xie and V.V. Phoha. Web user clustering from access log using belief function. In *Proceedings of the 1st international conference on Knowledge capture*, pages 202–208. ACM, 2001.

[37] R.R. Yager and L. Liu. *Classic works on the Dempster-Shafer theory of belief functions*. Springer Verlag, 2008.

[38] Osmar R. Zaïane and Maria-Luiza Antonie. Classifying text documents by associating terms with text categories. In *Proceedings of the 13th Australasian database conference - Volume 5*, ADC '02, pages 215–222, 2002.

[39] N. Zhong, Y. Li, and S.T. Wu. Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, DOI: http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.211.

[40] X. Zhou, S.T. Wu, Y. Li andY. Xu, R.Y.K. Lau, and P. Bruza. Utilizing search intent in topic ontology-based user profile for web mining. In *Proceeding of 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 558–561, 2006.

Analysing Effect of Database Grouping on Multi-Database Mining

Animesh Adhikari, Lakhmi C. Jain, Sheela Ramanna

Abstract — **In many applications we need to synthesize global patterns in multiple large databases, where the applications are independent of the characteristics of local patterns. Pipelined feedback technique *(*PFT*)* seems to be the most effective technique under the approach of local pattern analysis *(*LPA*)*. The goal of this paper is to analyse the effect of database grouping on multi-database mining. For this purpose we design a database grouping algorithm. We introduce an approach of non-local pattern analysis *(*NLPA*)* by combining database grouping algorithm and pipelined feedback technique for multi-database mining. We propose to judge the effectiveness of non-local pattern analysis for multi-database mining. We conduct experiments on both real and synthetic databases. Experimental results show that the approach to non-local pattern analysis does not always improve the accuracy of mining global patterns in multiple databases.**

Index Terms — **Local pattern analysis, Multi-database mining, Non-local pattern analysis, Pipelined feedback technique, Synthesis of patterns**

I. Introduction

MULTI-database mining is strategically an essential area of data mining. This is because of the fact that in many applications we need to process data from various sources [12], [13], [4], [8]. As a result, research in multi-database mining is gaining momentum [18], [20], [6].

In many situations data are collected from different regions across the globe. It might be possible to move data from one place to another place for some applications that are independent of the local properties of databases. The goal of this paper is to judge whether one could improve mining global patterns by sacrificing local properties of patterns in multi-databases. In an earlier work [7], we have shown that PFT improves the quality of global patterns significantly as compared to an existing technique [15], [17], [19], [5] that scans each database only once. In an effort to make further improvements, we introduce non-local pattern analysis for multi-database mining and propose to study its effectiveness in synthesizing global patterns. There are two primary reasons for non-local pattern analysis (i) the local properties of patterns need not always to be preserved; (ii) the number of estimations of a pattern might get decreased.

Local pattern analysis [19], [5] is an important approach of mining multiple large databases. One could obtain reasonably good solutions for a large class of problems. In local pattern analysis, each local database is mined locally. Then every branch forwards the local pattern base to the central location. All the pattern bases are then processed for synthesizing global patterns in multiple databases. It is important to observe that the same pattern might not get reported from every local database. As a result, the local pattern analysis is an approximate method of mining multiple large databases. If we are able to amalgamate all the databases together then there is no difference between mono-database mining and multi-database mining. There might be different reasons in different contexts that prohibit us to amalgamate all the databases together [6]. The next question comes to our mind is that whether one could reduce the frequency of database mining. In this regard, there are two extreme cases of multi-database mining viz., mono-database mining and local pattern analysis. Mono-database mining is used when there is a possibility of clubbing all the local databases. But the latter is used when each local database requires mining locally. In the first case, the frequency of mining database is one. But the frequency of mining is equal to the number of local databases in case of local pattern analysis. In view of reducing the frequency of mining, one may need to group the databases and then each group of databases is mined separately. Moreover, when we group some databases, the databases in a group are mined together. Thus, the number of estimations of a pattern will be reduced. For the purpose of constructing groups we consider that the groups of databases are mutually exclusive and exhaustive. The mutually exclusiveness property ensures that a database belongs to only one of the different groups. On the other hand exhaustiveness property ensures that each database belongs to a group. We club all the databases in a group for the purpose of multi-database mining. In this arrangement one needs to estimate a pattern less number of times, but local properties of a pattern may not get restored. This may have a bearing on the quality of the global patterns. In this paper, we investigate whether such an arrangement of local databases enhances accuracy of the global patterns.

Grouping of databases seems to an important issue for discovering knowledge in multiple databases. Wu and Zhang.

Animesh Adhikari is with the Department of Computer Science, S P Chowgule College, Goa, India (phone: 91-0832-2759504; fax: 91-0832-2759067; e-mail: animeshadhikari@yahoo.com).

Lakhmi C. Jain is with the School of Electrical and Information Engineering, University of South Australia, Mawson Lakes Campus, Australia (e-mail: Lakhmi.Jain@unisa.edu.au).

Sheela Ramanna is with the Department of Applied Computer Science, University of Winnipeg, Winnipeg, Canada (e-mail: s.ramanna@uwinnipeg.ca).

[16] have proposed a similarity measure sim_1 to identify similar databases based on item similarity. The authors have designed a clustering algorithm based on measure sim_1 to cluster databases for the purpose of selecting relevant databases. Such clustering is useful when the similarity is based on items in different databases. Item similarity measure sim_1 might not be useful in many multi-database mining applications where clustering of databases is based on some other criteria. For example, if we are interested in the databases based on transaction similarity then the above measures might not be appropriate. We have designed an algorithm for database clustering based on transaction similarity [4]. For this purpose, we have proposed a similarity measure $simi_1$ to cluster databases. One could group some objects based on an external criterion also. For example, the available main memory could pose a constraint in multi-database mining. It might be difficult to mine all the databases together when the databases are large. We will discuss later how the available main memory induces database grouping for the purpose of multi-database mining.

In an earlier work [7], we performed many experiments using different multi-database mining techniques (MDMTs). Experimental results have shown that PFT outperforms each of the existing techniques that scans a database only once. We introduce an approach of non-local pattern analysis based on PFT. For the purpose of completeness we present PFT in Section III.

Data mining applications based on multiple databases could be broadly categorized into two groups. The applications in the first group are based on patterns in individual databases. On the other hand, the second group of applications deals with the global patterns in multiple databases that are distributed in different geographical regions. Our goal is to study the effectiveness of non-local pattern analysis for mining global patterns in multiple databases. In many applications one may not have any restriction on moving a local database from one branch to another branch. Therefore, one could amalgamate a few branch databases and then mine a group of databases together. Then another group of databases could be formed and mined together, and so on. Finally, one could synthesize global patterns from the patterns in these groups of databases. We propose to study the effect of such grouping on synthesizing global patterns.

Rest of the paper is organized as follows. In Section II, we discuss related work. We present pipelined feedback technique in Section III. In Section IV, we introduce a non-local pattern analysis. We present a heuristic-based grouping algorithm in support of non-local pattern analysis. A discussion on finding the best grouping can be found in Section V. We present experimental results in Section VI.

II. Related Work

Zhang et al. [17] have proposed algorithm *IdentifyExPattern* for identifying global exceptional patterns in multi-databases. Here every local database is mined separately at random order using mono-database mining technique for synthesizing global exceptional patterns. As a result, the synthesized global patterns might deviate significantly from the true global patterns. We have proposed an algorithm *Association-Rule-Synthesis* [5] for synthesizing association rules in multiple real databases. This algorithm is useful for real databases, where the trend of the customers' behaviour exhibited in one database is usually present in other databases. For synthesizing high frequency association rules, Wu and Zhang [15] have proposed *RuleSynthesizing* algorithm for synthesizing high frequency association rules in multiple databases. Based on the association rules in different databases, the authors have estimated weights of different databases. Let w_i be the weight of the i-th database, $i = 1, 2, \ldots, n$. Without any loss of generality, let the association rule r be extracted from the first m databases, for $1 \leq m \leq n$. Actual support of r in D_i, $supp_a(r, D_i)$, has been assumed as 0, for $i = m + 1, m + 2, \ldots, n$. Then the support of r in D has been synthesized as follows.

$$supp_s(r, D) = w_1 \times supp_a(r, D_1) + \ldots + w_m \times supp_a(r, D_m) \qquad (1)$$

This method is an indirect approach and computationally expensive as compared to other techniques. Existing parallel mining techniques [2], [9] could also be used to deal with multiple large databases. In the context of pattern synthesis, Viswanath et al. [14] have proposed a novel pattern synthesizing method called *partition based pattern synthesis* which can generate an artificial training set of exponential order when compared with that of the given original training set.

III. Pipelined Feedback Technique (PFT)

For the purpose of completeness, we first present an overview of PFT [7]. Consider a multi-branch organization that collects data from multiple local branches. Let D_i be the database corresponding to the i-th branch, $i = 1, 2, \ldots, n$. Also let LPB_i be the local pattern base for D_i, $i = 1, 2, \ldots, n$. Also, let D be the union of all branch databases.

Let $D_1, D_2, \ldots, D_n$ be an arrangement of mining databases. First D_1 is mined using a mono-database mining technique [3], [11], and local pattern base LPB_1 is extracted. While mining D_2, all the patterns in LPB_1 are extracted irrespective of their values of interestingness measures such as minimum support and minimum confidence. Apart from these patterns, some new patterns that satisfy user-defined thresholds of interestingness are also extracted. In general, while mining D_i all the patterns in D_{i-1} are extracted irrespective of their values of interestingness, and some new patterns that satisfy user-defined thresholds of interestingness are also extracted. Due to this nature of mining each database, the technique is called a feedback model. Thus, $|LPB_{i-1}| \leq |LPB_i|$, $i = 2, 3, \ldots, n$. There are $n!$ arrangements of pipelining for n databases. All the arrangements of databases might not produce the same mining result. If the number of local patterns increases, we get more accurate global patterns and a better analysis of local patterns. An arrangement of local databases would produce near optimal result if $|LPB_n|$ is maximal. Let $size(D_i)$ be the size of D_i (in bytes), $i = 1, 2, \ldots, n$. We shall follow the following rule of thumb regarding the arrangements of databases for the purpose of mining: The number of patterns in D_{i-1} is greater than or equal to the number of patterns in D_i, if $size(D_{i-1}) \geq size(D_i)$, $i = 2, 3, \ldots, n$. For the purpose of increasing number of local patterns, D_{i-1} precedes D_i in the pipelined arrangement of mining databases if $size(D_{i-1}) \geq size(D_i)$, $i = 2, 3, \ldots, n$. Finally, we analyze the patterns in LPB_1, LPB_2, …, LPB_n for synthesizing global patterns, or analyzing local patterns.

Most of the databases are sparse. A pattern might not get reported from all the databases. However, once a pattern gets mined from a database, it also gets reported from the remaining databases in the pipeline. Thus, PFT improves the accuracy of multi-database mining. In the Section IV, we shall introduce an approach of non-local pattern analysis and we analyse its effectiveness in Section VI.

For synthesizing global patterns in D we discuss here a simple pattern synthesizing (SPS) algorithm with the help of itemset pattern in a database. Without any loss of generality, let the itemset X be extracted from the first m ($\leq n$) databases. Then synthesized support of X in D could be obtained as follows:

$$supp_s(X, D) = \frac{1}{\sum_{i=1}^{n} | D_i |} \times \sum_{i=1}^{m} [\, supp_a(X, D_i) \times | D_i | \,] \quad (2)$$

The accuracy of global pattern X increase as m approaches to n. The concepts of accuracy and error of a pattern are opposite to each other. When the error of a pattern increases, we say that its accuracy decreases, and vice-versa. We explain the concept of error in the following section.

A. Error

Let $D_1, D_2, \ldots, D_n$ be n branch databases. Also, let $size(D_1) \geq size(D_2) \geq \ \ldots \geq size(D_n)$. In PFT, the databases are mined according to the following order: $D_1, D_2, \ldots, D_n$. An itemset X gets reported from some of the given databases. In PFT, once X is reported from one of the given databases, then it also gets mined from the remaining databases. Suppose X is reported first time from D_k at minimum support level α, for $1 \leq k \leq n$. Then the error of mining X in D could be expressed as follows:

$$\text{Error}\,(X, D) = |supp_a(X, D) - supp_e(X, D)| \quad (3)$$

where, $supp_a(X, D)$ and $supp_e(X, D)$ denote the actual (apriori) support [3] and the estimated support of X in D, respectively. The supports $supp_a(X, \bigcup_{i=k+1}^{n} D_i)$ and $supp_e(X, \bigcup_{i=k+1}^{n} D_i)$ are the same, since X gets reported from the databases D_{k+1}, D_{k+2}, …, and D_n at minimum support level α. Thus, the error of mining X in D could be expressed as follows:

$$\text{Error}(X, D) = \left| supp_a(X, \cup_{i=1}^{k} D_i) - supp_e(X, \cup_{i=1}^{k} D_i) \right| \quad (4)$$

As the value of k increases, the amount of error increases provided the method of estimating support remains the same. Therefore, if the itemset X gets mined early in the pipelined arrangement, then amount of error decreases. In other words, as the number of estimations reduces, the error of mining itemset X reduces. This is an important observation and has been applied to the proposed non-local pattern analysis. Let S be the set of all itemsets synthesized from D. Then the average error (AE) of the experiment could be defined as follows:

$$\text{AE} = \frac{1}{|S|} \sum_{X \in S} Error(X, D) \quad (5)$$

Also, one could define maximum error (ME) of the experiment as follows:

$$\text{ME} = maximum\ \{Error\ (X, D) |\ X \in S\} \quad (6)$$

IV. Non-Local Pattern Analysis (NLPA)

Consider a multi-branch organization that has n ($\geq$ 2) branches. Suppose that each branch maintains a database of all local transactions. The goal of this paper is to investigate whether one could improve multi-database mining by sacrificing local properties of the patterns. In view of this one could group databases induced by available main memory. Let k be the number of groups of databases. Different groups of databases are given as follows: $\{D_{11}, D_{12}, \ldots, D_{1n_1}\}$, $\{D_{21}, D_{22}, \ldots, D_{2n_2}\}$, …, $\{D_{k1}, D_{k2}, \ldots, D_{kn_k}\}$, where $D_{ij} \in \{D_1, D_2, \ldots, D_n\}$, for $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$; $\sum_{i=1}^{k} n_i = n$; $n_i \geq 1$. Afterwards each group of databases are amalgamated and mined. The crux of non-local pattern analysis is how to group the databases so that one could mine each group of databases effectively within the limited memory. We formulate the problem of grouping databases as follows.

A. Grouping Databases

Multi-database mining could be performed by amalgamating some local databases and mining them together. But the performance of data mining process seems to be constrained by size of the main memory. If the available main memory is less, it might take a longer time to accomplish the mining task. During the grouping process, we shall continue to club databases as long as main memory is available. Let β be the optimum size of available main memory. Let $size(D)$ be the size of database D. Then the problem of grouping databases can be stated as follows:

We are given a set of numbers S = {*size*(D_1), *size* (D_2), …, *size* (D_n)}. *Our objective is to find r subgroups* S_1, S_2, … S_r, *for some* $r \geq 1$, *so that* $\sum_{i=1}^{r-1}\left(\beta - \sum_{j=1}^{n_i} size\left(D_{ij}\right)\right)$ *is a minimal, where the following conditions are true.*

(i) $\sum_{j=1}^{n_i} size\,(D_{ij}) \leq \beta$, *for i = 1, 2, …, r, and* $D_{ij} \in \{D_1, D_2, \ldots, D_n\}$

(ii) S_i = {*size*(D_{i1}), *size*(D_{i2}), …, *size*(D_{in_i})}, *and* $S_i \subseteq S$, *i = 1, 2, …, r*

(iii) $S_i \cap S_j = \phi$, $\forall$ *i ≠j, and* $\bigcup_{i=1}^{r} S_i = S$

$\beta - \sum_{j=1}^{n_i} size\left(D_{ij}\right)$ is the amount of *unutilized space* for the i-th group, $i = 1, 2, \ldots, r$. The goal of the grouping process is to reduce the total amount of unutilized spaces. In the next section, we propose a heuristic algorithm that utilizes main memory effectively.

B. An Heuristic Algorithm for Grouping Databases

As the number of groups decreases, one needs to estimate a global pattern fewer number of times. If the number of groups is one then all the patterns are exact and become true representative of the multiple databases. Given a limited amount of memory, it is important to group the databases so that it can fit best in the main memory. During the grouping process, if the larger databases are not considered at the early stage of grouping, then it could pose problems. As a result, the number of groups might increase. Smaller databases can be accommodated in a group easily, since their sizes are small. We apply this heuristic to design a grouping algorithm. Let us take an example to illustrate the grouping process.

Example 1. Let N be the set of sizes of the given databases. Let N be {139, 29, 43, 152, 165, 74, 5, 120}. Also let β be 200. First we sort the numbers in N in non-decreasing order. The ordered numbers are given as follows: 5, 29, 43, 74, 120, 139, 152, 165. The maximum size among the given databases is 165 bytes. First, we form a group with the database of size 165 bytes. Otherwise, it might cause producing a larger amount of unutilized space. Then along with the database of size 165 bytes, we club the database of size 29 bytes so that their sum 194 still remains less than or equal to 200. The database of size 29 bytes is obtained by searching the list from the right hand side. Any database of size in between 29 bytes and 165 bytes can not be clubbed with the database of size 165 bytes, since their sum would exceed 200 bytes. No more databases can be clubbed with them. Otherwise, their sum could exceed the available memory. In this case, the available memory is 200 bytes. As a result, the first group G_1 = {165, 29} is formed with an unutilized space of 6 bytes. Now we consider the database of size 152 bytes, since it is the second maximum among the given database sizes. Proceeding in the same way, one could form the second group as G_2 = {152, 43, 5} with an unutilized space of 0 byte. Then the next group G_3 = {139} is formed with unutilized space of size 61 bytes. The final group is G_4 = {120, 74} with unutilized space of 6 bytes. The total amount of unutilized spaces is equal to (6 + 0 + 61 + 6) bytes i.e., 73 bytes. Such grouping of databases might not be unique. For example, there exists another grouping of databases viz., {{152, 43, 5}, {165}, {139, 29}, {120, 74}}, that results in the same amount of unutilized spaces. •

Lemma 1. Let β be the optimum size of available main memory. Also, let D_{ij} be a database in group G_i, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, r$. Then the following grouping results in the same amount of unutilized spaces, provided $|G_j| + |D_{ik}| \le \beta$: $G_1, G_2, \ldots, G_{i-1}, G_i - \{D_{ik}\}, G_{i+1}, \ldots, G_{j-1}, G_j \cup \{D_{ik}\}, G_{j+1}, \ldots, G_r$, for some $i \neq j$. •

Based on the procedure illustrated in Example 1, we present here a heuristic algorithm, *Database-Grouping*, as follows.

```
procedure Database-Grouping ($n$, $A$, $\beta$)
Input:
n: number of databases
A: array of database sizes
$\beta$: maximum available memory (in bytes)
Output:
k: number of groups
G: two dimensional array representing different groups
01: sort A in non-decreasing order;
02: let k = 0;
03: for i = 1 to n do
04:  allocation(i) = 0;
05: end for
06: let index = n;
07: while (index $\ge$ 1) do
08:  let i = index; let sum = 0; let col = 1;
09:  increment k by 1;
10:  while (sum $\le \beta$) and (i $\ge$ 1) do
11:   if (sum + A(i) $\le \beta$) and (allocation(i) = 0) then
12:    sum = sum + A(i); allocation(i) = 1;
13:    increment col by 1; G(k, col) = A(i);
14:   end if
15:   decrease i by 1;
16:  end while
17:  G(k, 1) = col-1;
18:  let j = n;
19:  while (allocation(j) $\neq$ 0) and (j $\ge$ 1) do
20:   decrement j by 1;
21:  end while
22:  let index = j;
23: end while
24: for i = 1 to k do
25:  display the members of the i-th group;
26: end for
end procedure
```

We explain here the different variables and parts of the above algorithm. The number of groups is returned through the variable k. Here G is a two dimensional matrix that stores the output groups. The i-th row of G stores the i-th output group, $i = 1, 2, \ldots, k$. The first element of each row contains the number of elements in that group as noted in line 17. The subsequent elements are the database sizes in that group. The databases, whose sizes are kept in a group, are required to be clubbed for the purpose of mining. Initially, all the databases are unallocated (lines 03-05), since there exists no group. The database having a maximal size is allocated first in a group. Therefore, *index* variable gets initialized to n (line 06). The inner *while-loop* constructs a group of databases that are amalgamated afterwards for the purpose mining (lines 10-16). When a database is included in a group, the corresponding allocation tag is changed to 1 (line 12). Lines 18-22 help finding the next position (*index*) in the array A from which we start allocating the element for the next group. All the elements at the right side of current value of *index* are allocated to different groups.

Algorithm *Database-Grouping* forms k groups from the given databases, for some $k \le n$. Once the groups are formed, then we amalgamate the databases in each group for the purpose of mining. Accordingly, we have k amalgamated databases. We then follow pipelined feedback technique for mining these k databases.

The accuracy of synthesized patterns would depend on the sizes of the databases. In PFT, we mine first the database having the maximum size. It is expected that the database having the largest size would produce the maximum number of patterns. Further, PFT extracts all the previously extracted patterns irrespective of their interestingness values. Thus, it is always better to mine the largest database right at the beginning. The procedure *Database-Grouping* helps maximizing the database at every step by clubbing the databases. Moreover, it applies a heuristic approach while forming a group of databases.

In the context of mining time-stamped databases [8], *Database-Grouping* algorithm might play an important role. The time granularity of time-stamped databases is an important issue. Again, the time granularity would depend on an application. If time granularity is smaller, for example a month, then each of the monthly databases is expected to smaller. The procedure *Database-Grouping* would produce better grouping of databases. As a result, *Database-Grouping* algorithm is expected to produce good grouping when the size of each database is small.

Lemma 2. Let n be number of databases and k be the number of groups returned by the *Database-Grouping* algorithm. The time complexity of the algorithm is $O(n \times k)$.

Proof. *For-loop* in lines 3-5 takes $O(n)$ time. The algorithm returns k groups. Therefore, the outer *while-loop* in lines 7-23 repeats k times. The inner *while-loop* in lines 10-16 could repeat $O(n)$ times for each iteration of outer *while-loop*. Also, the *while-loop* in lines 19-21 could repeat n times for an iteration of outer *while-loop*. In lines 26-28, we display all the members in every group. Therefore, it takes $O(n)$ time. Thus, the time complexity of the algorithm is *maximum* $\{O(n), O(n \times k), O(n)\}$, i.e., $O(n \times k)$ time. •

C. Accuracy of mined patterns

If all the branch databases are amalgamated and mined then there is no difference between multi-database mining and mono-database mining. In this case a reported pattern is 100% accurate. But such situation may not exist always. Many branch databases could be very large. As a result the data mining process could consume unreasonable amount of time. In some cases it might not be possible to complete the data mining task. As a result a multi-database mining technique might report approximate patterns. An approximate pattern is not true representative pattern in multiple databases.

In our earlier work [7], we have noted that the accuracy of a mined pattern using PFT is generally higher than that of any other existing technique. This is true because of the fact that once a pattern is reported from a branch database, it also gets reported from the databases mined afterwards. If we can increase the size of each group (G_i) as much as possible by amalgamating branch some databases (D_j), the experimental results have shown that the average accuracy of a mined pattern might not decrease, for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, n$. As we increase the size of each G_i, we expect more patterns to be generated at each stage. Specifically, if a large number of patterns are reported at the initial stages of mining then the accuracy of those patterns, when synthesized globally, become higher. This is because of the fact that if a pattern is reported at any stage then it also gets reported subsequently due the application of feedback mechanism. Let us consider those patterns that are reported at the latter stages. These patterns might differ significantly from the actual global patterns. Therefore, the error of the experiment, AE and / or ME, might be more for non-local pattern analysis that that of PFT.

It might be appealing if one attaches depth of data mining with a mined pattern. We define *depth* of a pattern in multi-database mining as the fraction of total sizes of group databases from which a pattern gets extracted to the total size of all databases. Let $G_1, G_2, \ldots, G_r$ be the group of databases mined sequentially. Let pattern p be reported first time from the k-th group i.e., G_k. If p is an itemset pattern, then one could report its *depth* along with its support [1]. Thus,

$$depth\,(p) = (|G_k| + |G_{k+1}| + \ldots + |G_r|) \,/\, |D|,$$

where $(|G_1| + |G_2| + \ldots + |G_r|) = |D|$, $0 < depth\,(p) \leq 1$. Depth of a pattern represents the amount of data from which it has been extracted from a multi-database environment. If the depth of p is 1, then it is exact. One could discard a pattern if its depth is low.

V. An Optimal Grouping of Databases

Let us refer to algorithm *Database-Grouping* presented in Section IV. In most of the cases, it produces good grouping of databases. But it may not result in an optimal grouping for the purpose of multi-database mining. One could determine all possible groupings of databases at a given a set of databases and β. Then one could find the amount of unutilized spaces for every grouping. In the worst case one needs $O(n^2)$ comparisons to form a group, where n is the number of databases. Thus, the worst case complexity of such optimal algorithm is $O(n^2 \times k)$, where k is the number of groups. Such an algorithm might not be always attractive when a simpler algorithm like *Database-Grouping* produces an optimal result in the most of cases.

VI. Experimental Results

We have carried out several experiments to study the proposed approach of mining global patterns in multiple large databases. All the experiments have been implemented on a 2.8 GHz Pentium D dual core processor with 988 MB of memory using visual C++ (version 6.0) software. We present experimental results using synthetic dataset *T10I4D100K* [10] and two real datasets *retail* [10] and *BMS-Web-Wiew-1* [10]. We present some characteristics of these datasets in Table I. Let *NT*, *AFI*, *ALT*, and *NI* be the number of transactions, average frequency of an item, average length of a transaction, and number of items in a database, respectively. Each of the above datasets is divided into 10 databases for purpose of conducting our experiments.

TABLE I
DATASET CHARACTERISTICS

Dataset	*NT*	*ALT*	*AFT*	*NI*
T10I4D100K	1,00,000	11.10	1276.12	870
retail	88,162	11.31	99.67	10,000
BMS-Web-Wiew-1	1,49,639	2.00	155.71	1,922

TABLE II
T10I4D100K DATABASE CHARACTERISTICS

DB	NT	*size*(DB)	ALT	AFI	NI
T0	1,000	40	11.09	12.70	795
T1	2,000	81	11.18	24.43	834
T2	3,000	119	11.01	35.45	847
T3	4,000	159	11.01	46.84	855
T4	8,000	323	11.15	93.79	866
T5	10,000	400	11.05	115.93	867
T6	12,000	483	11.12	140.28	866
T7	15,000	605	11.13	175.24	867
T8	20,000	807	11.14	233.31	869
T9	25,000	1,027	11.07	290.38	867

TABLE III
retail DATABASE CHARACTERISTICS

DB	NT	*size*(DB)	ALT	AFI	NI
R0	1,000	36	9.52	5.11	1,000
R1	2,000	96	11.91	11.57	830
R2	3,000	143	11.72	16.22	862
R3	4,000	181	11.17	20.15	873
R4	8,000	358	11.10	4015	899
R5	10,000	473	11.49	49.82	1,097
R6	12,000	565	11.33	55.91	1,218
R7	14,000	634	10.76	58.34	1,311
R8	16,000	744	11.04	65.80	1,389
R9	18,162	922	11.89	77.87	1,500

TABLE IV
BMS-Web-Wiew-1 DATABASE CHARACTERISTICS

DB	NT	*size*(DB)	ALT	AFI	NT
B0	1,000	10	2.0	5.13	195
B1	2,000	22	2.0	3.56	157
B2	3,000	35	2.0	5.01	77
B3	5,000	63	2.0	3.05	1637
B4	10,000	131	2.0	6.23	1605
B5	15,000	205	2.0	1500	10
B6	20,000	273	2.0	2000	10
B7	25,000	341	2.0	2500	10
B8	30,000	410	2.0	3000	10
B9	38,639	528	2.0	3863.8	10

We have generated these databases arbitrarily consisting of a good mix of small and large databases. The databases obtained from *T10I4D100K*, *retail* and *BMS-Web-Wiew-1* are named as *Ti*, *Ri*, and *Bi* respectively, for *i* = 0, 1, …, 9 and subsequently referred to as input databases. Some characteristics of these input databases are presented in Tables II, III, and IV. Let N_{T} be {40, 81, 119, 159, 323, 400, 483, 605, 807, 1,027}, the set of sizes of databases obtained from *T10I4D100K*. Let N_{R} be {36, 96, 143, 181, 358, 47*3*, 565, 634, 744, 922}, the set of sizes of databases obtained from *retail*. Also, let N_{B} be {10, 22, 35, 63, 131, 205, 273, 341, 410, 528}, the set of sizes of databases obtained from *BMS-Web-Wiew-1*. In Table V, we present some outputs showing that the proposed non-local pattern analysis does not always improve accuracy of patterns in multiple large databases.

TABLE V
ERROR OF THE EXPERIMENTS AT A GIVEN MINIMUM SUPPORT

Dataset	*T10I4D100K*	*retail*	*BMS-Web-Wiew-1*
Minimum support	0.045	0.15	0.075
Error type	AE	AE	AE
MDMT: PFT + SPS	0.00451	0.00478	0.00206
MDMT: NLPA	0.00452	0.00499	0.00333
Error type	ME	ME	ME
MDMT: PFT + SPS	0.02411	0.01191	0.00702
MDMT: NLPA	0.02418	0.01270	0.00781

We apply *Database-Grouping* algorithm presented above. The choice of β for each of the three databases is an important issue. The sizes of *T10I4D100K*, *retail* and *BMS-Web-Wiew-1* are 3.83 MB, 3.97 MB, 1.97 MB respectively. Therefore, it might be possible to fit all the 10 databases in main memory for conducting experiments using each of the three datasets. But for the purpose of applying *Database-Grouping* algorithm one could consider β as little more than the maximum size of the generated databases, and accordingly, we taken β as 1,100 KB, 1,000 KB, and 700 KB for conducting experiments using datasets *T10I4D100K*, *retail* and *BMS-Web-Wiew-1*, respectively. The groups formed for above three datasets are given below:

Group corresponding to *T10I4D100K*, G_{T} = {{1027, 40}, {807, 159, 119}, {605, 483}, {400, 323, 81}} with the total amount of unutilized space is equal to (33 + 15 + 12 + 296) bytes i.e., 356 bytes.
Group corresponding to *retail*, G_{R} = {{922, 36}, {744, 181}, {634, 358}, {565, 143, 96}, {473}} with the total amount of unutilized space is equal to (42 + 75 + 8 + 196 + 527) bytes i.e., 848 bytes.
Group corresponding to *BMS-Web-Wiew-1*, G_{B} = {{528, 131, 35}, {410, 273, 10}, {341, 205, 63, 22}} with the total amount of unutilized space is equal to (6 + 7 + 69) bytes i.e., 82 bytes.

Now we club the databases in each group for purpose of mining multi-databases. Let the databases generated for the first, second and third groups be $D_{\mathrm{T}i}$, i = 1, 2, 3, 4; $D_{\mathrm{R}j}$, j = 1,2, 3, 4, 5; and $D_{\mathrm{T}k}$, k = 1, 2, 3, respectively. We present the databases after grouping in Tables VI, VII and VIII.

TABLE VI
NEW DATABASES GENERATED FROM *T10I4D100K*

Generated databases	Databases to be clubbed
D_{T1}	T_9, T_0
D_{T2}	T_8, T_3, T_2
D_{T3}	T_7, T_6
D_{T4}	T_5, T_4, T_1

TABLE VII
NEW DATABASES GENERATED FROM *RETAIL*

Generated databases	Databases to be clubbed
D_{R1}	R_9, R_0
D_{R2}	R_8, R_3
D_{R3}	R_7, R_4,
D_{R4}	R_6, R_2, R_1
D_{R5}	R_5

TABLE VIII
NEW DATABASES GENERATED FROM *BMS-WEB-VIEW-I*

Generated databases	Databases to be clubbed
D_{B1}	B_9, B_4, B_2
D_{B2}	B_8, B_6, B_0
D_{B3}	B_7, B_5, B_3, B_1

We have conducted experiments on the new databases by applying PFT and non-local pattern analysis. In Figs. 1, 2, and 3, we have presented results of AE with respect to minimum supports. Experimental results show that PFT reports more accurate global patterns than non-local pattern analysis in the most of the cases. Also, we observe that there no fixed trend of AE over the increased support values.

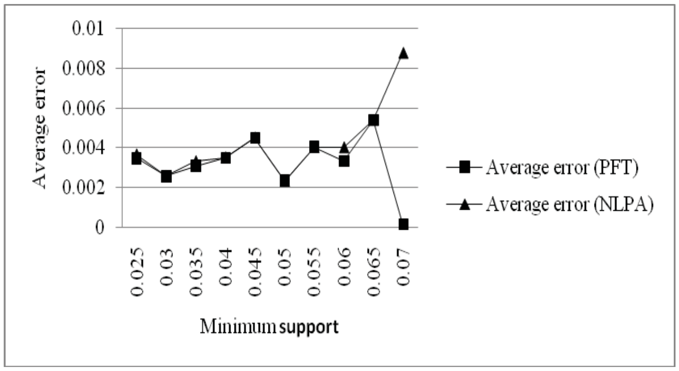


Fig. 1. Average error versus minimum support (for *T10I4D100K*)

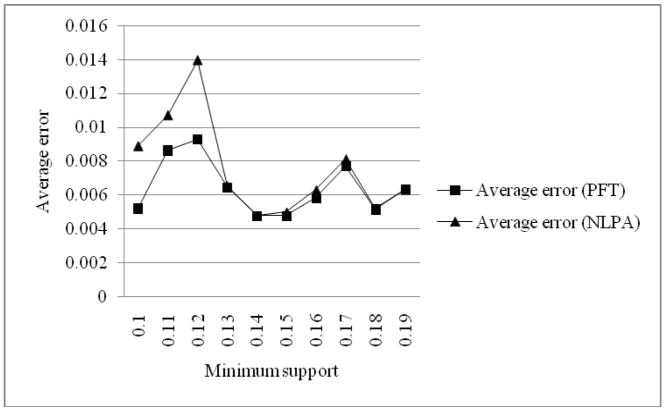


Fig. 2. Average error versus minimum support (for *retail*)

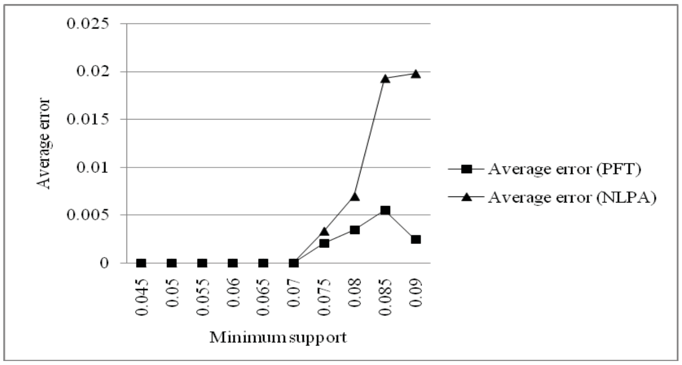


Fig. 3. Average error versus minimum support (for *BMS-Web-Wiew-1*)

VII. Conclusion

In this paper we have introduced non-local pattern analysis for multi-database mining in an attempt to study its effectiveness in synthesizing global patterns. A database grouping algorithm induced by main memory constraint has been introduced to applying non-local pattern analysis. Main memory constraint is an illustration of a criterion used for database grouping. Apparently non-local pattern analysis looks to be attractive, since the frequency of data mining is less as compared to local pattern analysis. As a result one needs to estimate a pattern lesser number of times for the purpose of synthesizing the global pattern. The drawback of non-local pattern analysis is that the patterns reported only from the last few groups might contribute significantly to the error of the experiment. This is due to the fact that a pattern is assumed absent when it does not get reported. Therefore, a mined pattern needs to be associated with the amount of data that it represents. For this purpose we have defined depth of a pattern in multi-database mining. A pattern becomes useless if its depth is low. We have conducted several experiments on real and synthetic datasets. Experimental results show that non-local pattern analysis might not be a better technique than PFT.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," In *Proceedings of ACM SIGMOD Conf. Management of Data*, 1993, pp. 207-216.

[2]R. Agrawal, and J. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engneering*, vol. 8, no. 6, pp. 962-969, 1999.

[3] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," In *Proceedings of VLDB*, 1994, pp. 487-499.

[4] A. Adhikari, and P. R. Rao, "Efficient clustering of databases induced by local patterns," *Decision Support Systems*, vol. 44, no. 4, pp. 925-943, 2008.

[5] A. Adhikari, and P. R. Rao, "Synthesizing heavy association rules from different real data sources," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 59-71, 2008.

[6] A. Adhikari, P. R. Rao, W. Pedrycz, *Developing multi-database mining applications*, Springer, 2010.

[7] A. Adhikari, P. R. Rao, B. Prasad, and J. Adhikari, "Mining multiple large data sources," *International Arab Journal of Information Technology*, vol. 7, no. 2, pp. 243-251, 2010.

[8] J. Adhikari, P. R. Rao, and A. Adhikari, “Clustering items in different data sources induced by stability,” *International Arab Journal of Information Technology*, vol. 6, no. 4, pp. 394-402, 2009.

[9] J. Chattratichat, J. Darlington, M. Ghanem, Y. Guo, H. Hüning, M. Köhler, J. Sutiwaraphun, H.W. To, and D. Yang, “Large scale data mining: Challenges, and responses,” In *Proceedings of KDD*, 1997, pp. 143-146.

[10] Frequent Itemset Mining Dataset Repository, http://fimi.cs.helsinki.fi/data/.

[11] J. Han, J. Pei, and Y. Yiwen, “Mining frequent patterns without candidate generation,” In *Proceedings of SIGMOD*, 2000, pp. 1-12.

[12] D. Page, and M. Craven, “Biological applications of multi-relational data mining,” *SIGKDD Explorations* vol. 5, no. 1, pp. 69-79, 2003.

[13] W. –C. Peng, Z. –X. Liao, “Mining sequential patterns across multiple sequence databases,” *Data & Knowledge Engineering*, vol. 68, no. 10, pp. 1014-1033, 2009.

[14] P. Viswanath, M.N. Murty, and S. Bhatnagar, 2006. “Partition based pattern synthesis technique with efficient algorithms for nearest neighbor classification,” *Pattern Recognition Letters*, vol. 27, no. 14, pp. 1714-1724, 2006.

[15] X. Wu, and S. Zhang, “Synthesizing high-frequency rules from different data sources,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 353-367, 2003.

[16] X. Wu, C. Zhang, and S. Zhang, “Database classification for multi-database mining,” *Information Systems*, vol. 30, no. 1, pp. 71-88, 2005.

[17] C. Zhang, M. Liu, W. Nie, and S. Zhang, “Identifying global exceptional patterns in multi-database mining,” *IEEE Computational Intelligence Bulletin*, vol 3, no 1, pp. 19-24, 2004.

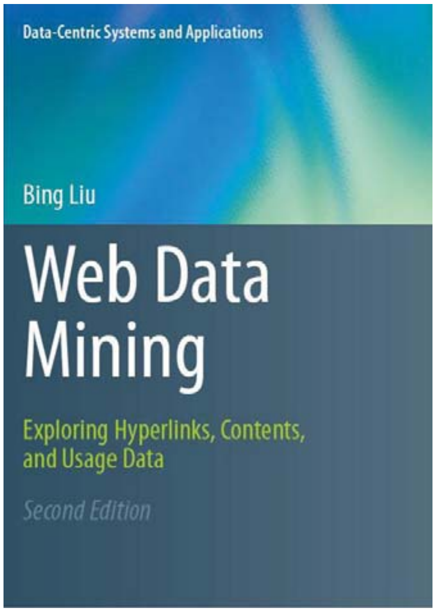
[18] S. Zhang, C. Zhang, X. Wu, *Knowledge discovery in multiple databases*. Springer, 2004.

[19] S. Zhang, X. Wu, C. Zhang, “Multi-database mining,” *IEEE Computational Intelligence Bulletin*, vol. 2, no. 1, pp. 5-13, 2003.

[20] S. Zhang, and M. J. Zaki, “Mining multiple data sources: Local pattern analysis,” *Data Mining and Knowledge Discovery* (*Special issue*), Springer, 2006.

Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

BY BING LIU– ISBN 978-3-642-19459-7


Data-Centric Systems and Applications
Bing Liu
Web Data Mining
Exploring Hyperlinks, Contents, and Usage Data
Second Edition

REVIEWED BY RUSSELL JOHNSON

Not so long ago only a small proportion of the information generated by organizations – that stored in structured data sources like databases and spreadsheets – was accessible for systematic computer-based search, classification and analysis. The techniques and algorithms of data mining were developed to extract useful patterns and knowledge from these structured sources. Today, the explosive growth of the World-Wide-Web, and a parallel development of private intranets, means that a much greater amount of information is potentially available for search and analysis, in a wider variety of formats and encompassing structured, semi-structured and unstructured data, from organized tables to multi-media clips. The established techniques of data mining have proved insufficient for this task.

Over the past decade, new techniques and algorithms have been developed which aim to discover useful information or knowledge from computer analysis of the hyperlink structures, page contents, and usage data of Web resources. *“Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data”* by University of Illinois Professor Bing Liu provides an in-depth treatment of this field.

In the introduction, Liu notes that to explore information mining on the Web, it is necessary to know data mining, which has been applied in many Web mining tasks. However, he points out that Web mining is not entirely an application of data mining. Due to the richness and diversity of information and other Web specific characteristics discussed above, Web mining has developed many of its own algorithms.

One of the standout features of Liu’s book is that it encompasses both data mining and Web mining. The first half of his book outlines the major aspects of data mining which Liu lists as supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining, and provides examples of how these techniques are used in Web mining.

In the second half, the author focuses on specific web mining techniques. Based on the primary kinds of data used in the mining process, Liu categorizes these into three types: Web structure mining, Web content mining and Web usage mining.

Web structure mining abstracts useful knowledge from the hyperlink structure of the Web, which search engines do to discover important Web pages. It is also possible to discover communities of users who share common interests. Liu points out that traditional data mining cannot perform such tasks because relational tables do not have link structures. *Web content mining* extracts useful information from the contents of Web pages, such as automatically classifying and clustering pages according to their topics. These tasks are similar to those in traditional data mining. However, additional tasks can be performed such as mining customer reviews and forum postings to determine customer opinions which are not traditional data mining tasks.

Web usage mining involves discovering user access patterns from Web usage logs, which record each user’s mouse clicks, and applying data mining algorithms to them.

Summary of Chapters

It is useful at this point to provide a summary of the chapter contents. As previously explained the book is divided into two main parts. Chapters 2-5 cover the major topics of data mining while the remaining chapters cover Web mining, including a chapter on Web search.

Chapter 1 provides a brief history of the web and of web mining. It also provides a useful summary of the contents of the book and how to read it.

Chapter 2 studies Association Rules and Sequential Patterns, which have been used in many Web mining tasks, especially in Web usage and content mining. Association rule mining finds sets of data items that occur together frequently. Sequential pattern mining finds sets of data items that occur together frequently in some sequences, and can be used to find regularities in the Web data.

Supervised learning (Chapter 3), or classification, is frequently used in both practical data mining and Web mining. It aims to learn a classification function (called a classifier) from data that are labeled with pre-defined classes or categories. The resulting classifier is then applied to classify future data instances into these classes.

In unsupervised learning (Chapter 4), the learning algorithm has to find the hidden structures or regularities in the data without any pre-defined classes, such as with clustering, which organizes data instances into groups or clusters according to their similarities (or differences).

Chapter 5 explains the Partially Supervised Learning model replacing the large number of manually-labeled examples required for the supervised learning model with a small set of labeled examples (data instances) and a large set of unlabeled examples for learning.

The section on web mining begins with Chapter 6 on Information Retrieval and Web Search. The vast scale of the web means that web search algorithms must not only be accurate but also efficient.

Chapter 7 (Social Network Analysis) outlines how hyperlinks are exploited for efficient Web search including looking at Google's hyperlink-based ranking algorithm, PageRank, which originated from social network analysis, and community finding algorithms which underpin social networking sites.

In Chapter 8 examines the techniques of web crawling – traversing the Web's hyperlink structure and locating pages linked by topic – which is usually the essential first step of Web mining or building a Web search engine

In Chapter 9 (Structured Data Extraction: Wrapper Generation) data mining techniques are explored for analyzing web pages based on structured data, possibly extracted from databases, to identify the underlying patterns and extract the data to provide value-added services such as comparative shopping.

Information Integration (Chapter 10) involves matching and integrating information from different web sources, especially structured data, to provide consistent and coherent database.

Chapter 11 (Opinion Mining and Sentiment Analysis) looks at techniques for analyzing and categorizing the huge amount of unstructured text on the Web to mine people's opinions and sentiments expressed in product reviews, forum discussions and blogs.

Finally, Chapter 12 on Web usage mining aims to study user clicks and their applications to e-commerce and business intelligence. The objective is to capture and model behavioral patterns and profiles of users who interact with a Web site to support *recommender* systems and other commercial applications.

Conclusion

Originally published in 2007, this revised and updated second edition is primarily intended as a textbook for an advanced computer science course. Each chapter has a reference section for future reading. In addition, a companion website at http://www.springer.com/3-540-37881-2 provides updates, lecture slides, implemented examples and other useful teaching resources.

Two chapters are mainly contributed by three other researchers Filippo Menczer,

Bamshad Mobasher, and Olfa Nasraoui. Menczer contributed Chapter 8 on Web crawling, while Bamshad and Olfa wrote most of Chapter 12 on Web usage mining.

The content of *"Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"* is comprehensive and in-depth. While the book's focus in web mining Liu recommends that students without a background in machine learning should not skip the sections on data mining. He also suggests that the early sections could provide the basis for an introductory data mining course, especially if integrated with the chapters on search engines and social networking which are of topical interest among students.

THE BOOK:

LIU, BING (2011) WEB DATA MINING, 2ND EDITION. 2011, XX, 622 P. ILLUS. SPRINGER.
ISBN: 978-3-642-19459-7

ABOUT THE REVIEWER:

RUSSELL JOHNSON
School of Engineering and Advanced Technology, Massey University, New Zealand. Contact him at: r.s.johnson@massey.ac.nz

Special issue on Advances in Web Intelligence

Stefan Rüger[a,**], Vijay Raghavan[b,*], Irwin King[c,*], Jimmy Xiangji Huang[d,*]

[a]*Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom*
[b]*The Center for Advanced Computer Studies, University of Louisiana at Lafayette, P.O. Box 44330, Lafayette, LA 70504-4330, USA*
[c]*Department of Computer Science & Engineering, 908 Ho Sing Hang Engineering Building, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, SAR*
[d]*School of Information Technology, York University, Toronto, Ontario, Canada M3J 1P3*

Abstract

We summarize the scientific papers of the special issue “Advances in Web Intelligence,”which will appear in the Neurocomputing journal, Volume 76, Issue 1 (2012)- published by Elsevier. These papers are substantially extended from original contributions to the Web Intelligence 2010 conference held in Toronto, Canada, in September 2010.

Keywords: special issue, web intelligence

1. Introduction

Web intelligence is commonly seen as a combination of applied artificial intelligence and information technology in the context of the web with a view to study and characterize emerging — or design new — products and services of the internet. The Web Intelligence conference, held every year between 2001 and 2010 with the exception of 2002, has recognized these new directions and provides a scientific forum for researchers and practitioners to further topics such as web intelligence foundations; world wide wisdom web (W4); web information retrieval and filtering; semantics and ontology engineering; web mining and farming; social networks and ubiquitous intelligence; knowledge grids and grid intelligence; web agents; web services; intelligent human-web interaction; web support systems; intelligent e-technology; and other related areas.

Web Intelligence 2010 received 313 submissions, of which 51 regular papers were accepted. The papers in this special issue reflect the trends at Web Intelligence 2010 and focus on particularly strong contributions to the conference, of which we invited 16 to the special issue. These authors since then expanded their Web Intelligence 2010 contribution significantly for the benefit of the Neurocomputing readership guided by an independent, critical peer review process that ultimately accepted 9 submissions. The areas to which the papers of this special issue contribute can broadly be characterized into content analysis (Section 2), social media and network analysis (Section 3) and machine learning for web intelligence (Section 4).

2. Content analysis

The paper “Multimodal Representation, Indexing, Automated Annotation and Retrieval of Image Collections via Nonnegative Matrix Factorization” by Caicedo et al. proposes a novel method of analyzing and generating multimodal image representations that integrate both visual features in images and their associated text information such as descriptions, comments, user ratings and tags. Experiments using Corel and Flickr data sets have demonstrated the advantage of nonnegative matrix factorization with asymmetric multimodal representation over other approaches such as direct matching and singular value decomposition.

Krestel and Fankhauser’s paper “Personalized Topic-Based Tag Recommendation” proposes an approach for personalized tag recommendation that combines tags derived via the probabilistic model of a web resource with those obtained from the user. The paper investigates simple language models as well as Latent Dirichlet Allocation as alternatives for modeling the resource content. Experiments on a real world dataset show that personalization improves tag recommendation, and the proposed approach significantly outperforms state-of-the-art approaches, such as FolkRank.

The next paper “On Optimization of Expertise Matching with Various Constraints” by Tang et al. studies mechanisms of assigning experts to a set of items such as submitted papers to a conference or to-be-reviewed products under constraints, for example, that the overall workload of individual experts is balanced and that each item has a certain number of reviews by senior and less senior experts. The paper formulates the expertise matching problem in a general constraint-based optimization framework that links the problem to a convex cost flow, which promises an optimal solution under various constraints. Tang et al. also propose an online matching algorithm that incorporates immediate user feedback. Experimental results validate the effectiveness of the proposed approach in the cases of reviewer to conference paper assignment and teacher to course assignment.

[*] Guest editor
[**] Managing guest editor

Reprinted from Neurocomputing, Vol 76, Issue 1 (2012), 48–49, Copyright (2012), with permission from Elsevier

3. Social media and network analysis

The paper “Characteristics of Information Diffusion in Blogs, in Relation to Information Source Type” by Kazama et al. introduces information diffusion properties to analyze the dynamics of blogs based on constructed subgraphs for information recommendation and ranking. The work focuses on three types of basic structures: information scattering, information gathering, and information transmission structures. With these information diffusion properties, the work is able to represent various social media characteristics and provide priority to different types of information sources.

“A Framework For Joint Community Detection Across Multiple Related Networks” by Comar et al. utilizes non-negative matrix factorization that combines information from multiple networks in order to identify communities and learn the correspondences among these networks simultaneously. The method has shown good performance over other approaches such as normalized cut and matrix factorization with experiments done on both synthetic as well as real-world wikipedia and digg data sets.

Largillier and Peyronnet demonstrate in their paper “Webspam Demotion: Low Complexity Node Aggregation Methods” a mechanism to lower the ranking of webspam, which are undesirable web pages that were created with the sole purpose of influencing link-based ranking algorithms to promote a particular target page. Webspam techniques evolve all the time, but almost inevitably they create a specific linking architecture around the target page to increase its rank. Largillier and Peyronnet study the effects of node aggregation of the well-known PageRank algorithm in presence of webspam. Their lightweight node aggregation methods aim to construct clusters of nodes that can be considered as a sole node in the PageRank computation. Experimental results show the promise of the presented webspam demotion approach.

4. Machine learning for web intelligence

Yan et al. propose in their paper “Semi-Supervised Dimensionality Reduction for Analyzing High-Dimensional Data with Constraints” a novel technique to address the problems of inefficient learning and costly computation in coping with high-dimensional data. The approach, termed Dual Subspace Projections, embeds high-dimensional data in an optimal low-dimensional space, which is learned with a few user-supplied constraints and the structure of input data. The method overcomes the model overfitting problem by simultaneously preserving both the structure of original high-dimensional data and user-specified constraints. Experiments on real datasets from multiple domains demonstrate that significant improvement in learning accuracy can be achieved via their dimensionality reduction technique, even with only a few user-supplied constraints.

Ramirez et al.’s paper “Topic Model Validation” considers the problem of performing external validation of the semantic coherence of topic models. Ramirez et al. generalize the Fowlkes-Mallows index, a clustering validation metric, for the case of overlapping partitions and multi-labeled collections rendering it suitable for assessing topic modeling algorithms. They also propose probabilistic metrics inspired by the concepts of recall and precision and show how these can be applied to validate and compare other soft and overlapping clustering algorithms.

In their paper “Modeling and Predicting the Popularity of Online Contents with Cox Proportional Hazard Regression Model”, Lee et al. propose a framework, which can be used for modeling and predicting the popularity of discussion forum threads based on initial observations of how the thread evolves and the number of comments. The underlying approach is rooted in survival analysis, which models the survival time until an event of a failure or death. Lee et al. model the lifetime of discussion threads and the number of comments that the contents receives, with a set of explanatory and externally observable factors, using the Cox proportional hazard regression model, which divides the distribution function of the popularity metric into two components: one which is explained by a set of observable factors, and another, a baseline survival distribution function, which integrates all the factors not taken into account. The methodology is validated with two datasets that were crawled from two different discussion fora.

Acknowledgements

The guest editors wish to thank the Editor-in-Chief Tom Heskes for giving us the opportunity to present the Advances in Web Intelligence; the reviewers, who have helped to select the very best papers from Web Intelligence 2010 and gave valuable feedback on the extended work in the papers of this special issue within a very short time-frame; the journal staff for a professional production; and, not least, the authors, who carry this field with much enthusiasm and scientific rigor.

References

Caicedo, J. C., BenAbdallah, J., Gonzalez, F. A., Nasraoui, O., 2012. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing 76 (1), 50–60.

Comar, P. M., Tan, P.-N., Jain, A. K., 2012. A framework for joint community detection across multiple related networks. Neurocomputing 76 (1), 93–104.

Kazama, K., Imada, M., Kashiwagi, K., 2012. Characteristics of information diffusion in blogs, in relation to information source type. Neurocomputing 76 (1), 84–92.

Krestel, R., Fankhauser, P., 2012. Personalized topic-based tag recommendation. Neurocomputing 76 (1), 61–70.

Largillier, T., Peyronnet, S., 2012. Webspam demotion: Low complexity node aggregation methods. Neurocomputing 76 (1), 105–113.

Lee, J. G., Moon, S., Salamatian, K., 2012. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. Neurocomputing 76 (1), 134–145.

Ramirez, E. H., Brena, R., Magatti, D., Stella, F., 2012. Topic model validation. Neurocomputing 76 (1), 125–133.

Tang, W., Tang, J., Lei, T., Tan, C., Gao, B., Li, T., 2012. On optimization of expertise matching with various constraints. Neurocomputing 76 (1), 71–83.

Yan, S., Bouaziz, S., Lee, D., Barlow, J., 2012. Semi-supervised dimensionality reduction for analyzing high-dimensional data with constraints. Neurocomputing 76 (1), 114–124.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

THE 2012 WORLD INTELLIGENCE CONGRESS

(WI+IAT+AMT+BI+ISMIS)

Macau, China
December 4- 7, 2012
http://degroup.sftw.umac.mo/~wic2012/
http://www.comp.hkbu.edu.hk/~wic/

World Intelligence Congress 2012 is being organized/sponsored by the Web Intelligence Consortium (WIC), the IEEE-CS Technical Committee on Intelligent Informatics (TCII), the IEEE-CIS Task Force on Brain Informatics and ACM SIGART as a special event of the Alan Turing Year (Centenary of Alan Turing's birth) in 2012. The congress includes five intelligent informatics related conferences co-located with the aim to facilitate interactions and idea exchange among researchers working on a variety of focused themes under a holistic vision for computing and intelligence in the post WWW era, and to promote and expedite new innovations for areas under intelligent informatics.

WI 2012
The 2012 IEEE/WIC/ACM International Conference on Web Intelligence

Web Intelligence (WI) explores the fundamental roles, interactions as well as practical impacts of Artificial Intelligence (AI) engineering and advanced information technology on the next generation of Web systems. Here AI-engineering is a general term that refers to a new area, slightly beyond traditional AI: brain informatics, human level AI, intelligent agents, social network intelligence and classical areas such as knowledge engineering, representation, planning, discovery and data mining are examples. Advanced information technology includes wireless networks, ubiquitous devices, social networks, and data/knowledge grids, as well as cloud computing, service oriented architecture.

IAT 2012
The 2012 IEEE/WIC/ACM International Conference on Intelligent Agent Technology

IAT 2012 will provide a leading international forum 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. By encouraging idea-sharing and discussions on the underlying logical, cognitive, physical, and sociological foundations as well as the enabling technologies of intelligent agents, IAT 2012 will foster the development of novel paradigms and advanced solutions in agent and multi-agent based computing.

AMT 2012
The 2012 International Conference on Active Media Technology

The rapid scientific and technological developments in human-centric, seamless interfaces, devices, connections, mobility, computing resources, computing environments and systems with their applications ranging from business and communication to entertainment and learning are collectively best characterized as Active Media Technology (AMT). AMT is a new area of intelligent information technology and computer science that emphasizes the proactive, seamless roles of interfaces, connections, and systems as well as new media in all aspects of digital life. An AMT based system offers active and transparent services to enable the rapid design, implementation and support of customized solutions. The first International Conference on Active Media Technology (AMT 2001) was held in Hong Kong in 2001. After 10 years went around the world, AMT 2012 will be held in conjunction with other 4 conferences in Macau where is very close to Hong Kong.

BI 2012
The 2012 International Conference on Brain Informatics

BI 2012 provides a leading international forum to bring together researchers and practitioners that explore the interplay between the studies of human brain and the research of informatics in diverse fields, such as computer science, information technology, artificial intelligence, Web intelligence, cognitive science, neuroscience, medical science, life science, economics, data mining, data and knowledge engineering, intelligent agent technology, human computer interaction, complex systems, and system science. On the one hand, studies on human brain model and characterize the functions of the human brain based on the notions of information processing systems. On the other hand, informatics-enabled brain studies, e.g., based on fMRI, EEG, MEG significantly broaden the spectrum of theories and models of brain sciences and offer new insights into the development of human-level intelligence. Web Intelligence centric information technologies are applied to support brain science studies. For instance, the wisdom Web and knowledge grids enable high-speed, large-scale analysis, simulation, and computation as well as new ways of sharing research data and scientific discoveries.

ISMIS 2012
The 20th International Symposium on Methodologies for Intelligent Systems

International Symposium on Methodologies for Intelligent Systems is an established and prestigious conference for exchanging the latest research results in building intelligent systems. Held twice every three years, the conference provides a medium for exchanging scientific research and technological achievements accomplished by the international community. The 20th International Symposium on Methodologies for Intelligent Systems - ISMIS 2012 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 (but are not limited to): active media, human-computer interaction, autonomic and evolutionary computation, digital libraries, intelligent agent technology, intelligent information retrieval, intelligent information systems, intelligent language processing, knowledge representation and integration, knowledge discovery and data mining, knowledge visualization, logic for artificial intelligence, music information retrieval, soft computing, text mining, web intelligence, web mining, web services, social computing, and recommender Systems.

ICDM 2012
The Twelfth IEEE International Conference on Data Mining
Brussels, Belgium
December 10-13, 2012
http://icdm2012.ua.ac.be/

The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications. In addition, ICDM draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing. By promoting novel, high quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state-of-the-art in data mining. Besides the technical program, the conference features workshops, tutorials, panels and, since 2007, the ICDM data mining contest.

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: data mining foundations, mining in emerging domains, methodological aspects and the KDD process, and integrated KDD applications, systems, and experiences. A detailed listing of specific topics can be found at the conference website.

BIBM 2011
IEEE International Conference on Bioinformatics & Biomedicine
Atlanta, Georgia, USA
November 12-15, 2011
http://www.cs.gsu.edu/BIBM2011/

IEEE BIBM 2011 will provide a general forum for disseminating the latest research in bioinformatics and biomedicine. It is a multidisciplinary conference that brings together academic and industrial scientists from computer science, biology, chemistry, medicine, mathematics and statistics. BIBM will exchange research results and address open issues in all aspects of bioinformatics and biomedicine and provide a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, simulation, ontology and other computational methods, as applied to life science problems, with emphasis on applications in high throughput data-rich areas in biology, biomedical engineering. IEEE BIBM 2011 intends to attract a balanced combination of computer scientists, biologists, biomedical engineers, chemist, data analyzer, statistician.

ICTAI 2012
The Twenty-fourth IEEE International Conference on Tools with Artificial Intelligence
Boca Raton，USA
October 1-3, 2012

The annual IEEE International Conference on Tools with Artificial Intelligence (ICTAI) provides a major international forum where the creation and exchange of ideas related to artificial intelligence are fostered among academia, industry, and government agencies. The conference facilitates the cross-fertilization of these ideas and promotes their transfer into practical tools, for developing intelligent systems and pursuing artificial intelligence applications. The ICTAI encompasses all technical aspects of specifying, developing and evaluating the theoretical underpinnings and applied mechanisms of the AI based components of computer tools (i.e. algorithms, architectures and languages).

Related Conferences

AAMAS 2012
The Eleventh International Conference on Autonomous Agents and Multi-Agent Systems
Valencia, Spain
June 4- 8, 2012
http://aamas2012.webs.upv.es/

The AAMAS conference series was initiated in 2002 in Bologna, Italy as a joint event comprising the 6th International Conference on Autonomous Agents (AA), the 5th International Conference on Multiagent Systems (ICMAS), and the 9th International Workshop on Agent Theories, Architectures, and Languages (ATAL). Subsequent AAMAS conferences have been held in Melbourne, Australia (July 2003), New York City, NY, USA (July 2004), Utrecht, The Netherlands (July 2005), Hakodate, Japan (May 2006), Honolulu, Hawaii, USA (May 2007), Estoril, Portugal (May 2008), Budapest, Hungary (May 2009), Toronto, Canada (May 2010), Taipei, Taiwan (May 2011). AAMAS 2012 will be held in June in Valencia, Spain.

AAMAS is the largest and most influential conference in the area of agents and multiagent systems, the aim of the conference is to bring together researchers and practitioners in all areas of agent technology and to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multiagent systems.

AAAI 2012
The Twenty-Sixth AAAI Conference on Artificial Intelligence
Toronto, Ontario, Canada
July 22-26, 2012
http://www.aaai.org/Conferences/AAAI/aaai12

The Twenty-Sixth Conference on Artificial Intelligence (AAAI 2012) will be held in will be held in Toronto, Ontario, Canada at the Sheraton Centre Toronto, from July 22–26, 2012. The purpose of the AAAI 2012 conference is to promote research in AI and scientific exchange among AI researchers, practitioners, scientists, and engineers in related disciplines. Details about the AAAI 2012 program will be published on http://www.aaai.org/Conferences/AAAI/aaai12 as they become available.

SDM 2012
The Twelfth SIAM International Conference on Data Mining

Anaheim, California, USA
April 26- 28, 2012
http://www.siam.org/meetings/sdm12/

Data mining is an important tool in science, engineering, industrial processes, healthcare, business, and medicine. The datasets in these fields are large, complex, and often noisy. Extracting knowledge requires the use of sophisticated, high-performance and principled analysis techniques and algorithms, based on sound theoretical and statistical foundations. These techniques in turn require powerful visualization technologies; implementations that must be carefully tuned for performance; software systems that are usable by scientists, engineers, and physicians as well as researchers; and infrastructures that support them.

This conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students and others new to the field to learn about cutting-edge research by hearing outstanding invited speakers and attending tutorials (included with conference registration). A set of focused workshops are also held on the last day of the conference. The proceedings of the conference are published in archival form, and are also made available on the SIAM web site.

IJCAI 2013
The Twenty-Third International Joint Conference on Artificial Intelligence

Beijing, China
August 5-9, 2013
http://ijcai-2013.org/

The Twenty-Third International Joint Conference on Artificial Intelligence IJCAI 2013 will be held in Beijing, China, August 5-9, 2013. Submissions are invited on significant, original, and previously unpublished research on all aspects of artificial intelligence. Details about the IJCAI 2013 program will be published on http://ijcai-2013.org/ as they become available.

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