# Introduction to the Domain-Driven Data Mining Special Section

Chengqi Zhang, *Senior Member*, *IEEE Computer Society*, Philip S. Yu, *Fellow*, *IEEE*, and David Bell

In the last decade, data mining has emerged as one of the most dynamic and lively areas in information technology. Although many algorithms and techniques for data mining have been proposed, they either focus on domain-independent techniques or on very specific domain problems. A general requirement in bridging the gap between academia and business is to cater to general domain-related issues surrounding real-life applications, such as constraints, organizational factors, domain expert knowledge, domain adaption, and operational knowledge. Unfortunately, these either have not been addressed, or have not been sufficiently addressed, in current data mining research and development.

By common consent, experience seems to indicate that real-world data mining must, in the majority of cases, consider and involve the domain experts' role, domain knowledge, business intelligence, human intelligence, network intelligence, social intelligence, domain-specific constraints, as well as organizational factors and social issues in practice. However, it is difficult to merge the above domain factors with data mining models and processes. It is also challenging to discover knowledge that will support users to take decision-making actions.

Domain-Driven Data Mining ($D^3M$) aims to develop general principles, methodologies, and techniques for modeling and merging comprehensive domain-related factors and synthesized ubiquitous intelligence surrounding problem domains with the data mining process, and discovering knowledge to support business decision-making.

This special section aims to report original, cutting-edge, and state-of-the-art progress in $D^3M$. It covers theoretical and applied contributions aiming to: 1) propose next-generation data mining frameworks and processes for actionable knowledge discovery, 2) investigate effective (automated, human&machine-centered and/or human-machined-co-operated) principles and approaches for acquiring, representing, modelling, and engaging ubiquitous intelligence in real-world data mining, and 3) develop workable and operational systems balancing technical significance and applications concerns, and converting and delivering actionable knowledge into operational applications rules to seamlessly engage application processes and systems.

This special section on $D^3M$ attracted 84 submissions. After a careful review process by international experts, seven papers were accepted that cover a wide research area in $D^3M$; one paper addresses the general framework of $D^3M$, four papers cover the approaches of $D^3M$, and two papers consider the workable applications of $D^3M$.

The first paper describes one possible framework of $D^3M$. In the paper, "Domain Driven Data Mining: Challenges and Prospects," Longbing Cao presents an overview of driving forces, theoretical frameworks, architectures, techniques, open issues, and case studies of $D^3M$. This is a comprehensive paper which formalizes $D^3M$ research to guide the future directions of this research area. The paper emphasizes the paradigm shift from "data-centered knowledge discovery" to "domain-driven actionable knowledge delivery." The main contribution of this paper is to shape the $D^3M$ area for current progress and future directions.

Four papers develop different approaches to $D^3M$. In the paper, "Bridging Domains Using World Wide Knowledge," Evan Wei Xiang, Bin Cao, Derek Hao Hu, and Qiang Yang design a new approach, called BIG (*Bridging Information Gap*), in which they introduce domain knowledge to help link the source and target data sets by introducing auxiliary data, even when the source and target data sets are far from each other. Using the auxiliary data, the authors can extract a "bridge" that allows the cross-domain text classification problem to be solved using standard semi-supervised learning algorithms. A main contribution of the paper is that previously untransferrable domains can now be successfully adapted for knowledge transfer.

In the paper, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," Claudia Marinica and Fabrice Guillet propose a new interactive approach to prune and filter discovered rules. The authors first propose using ontologies in order to improve user knowledge in the postprocessing task. Second, using ontology concepts, they introduce the Rule Schema formalism by extending the specification language for user expectations. An interactive framework is also designed to assist the user with the analyzing task. The main contribution of the paper is to overcome the drawbacks due the fact that "the usefulness of association rules is strongly limited by the huge amount of delivered rules." The authors have tested the approach on a real-life database, and the results have been validated by a domain expert.

Alex T.H. Sim, Maria Indrawan, Samar Zutshi, and Bala Srinivasan propose a framework to discover domain knowledge reporting it as coherent rules in "Logic-Based


- *C. Zhang is with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, Australia. E-mail: chengqi@it.uts.edu.au*
- *P.S. Yu is with the Department of Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607. E-mail: psyu@cs.uic.edu.*
- *D. Bell is with the School of Electronics, Electrical Engineering, and Computer Science, Queen's University Belfast, Belfast BT7 1NN, North Ireland, UK. E-mail: da.bell@qub.ac.uk.*


*For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org.*

Pattern Discovery." Coherent rules are discovered based on the properties of propositional logic and therefore require no background knowledge to generate them. The contribution of this paper is that association rules can be derived objectively and directly from the coherent rules discovered, without knowing the level of minimum support threshold required.

A novel active learning algorithm that poses generalized queries to the domain experts is proposed in the paper "Asking Generalized Queries to Domain Experts to Improve Learning" by Jun Du and Charles X. Ling. The power of such generalized queries is that one generalized query may be equivalent to many specific ones. However, overly general queries may receive highly uncertain answers from the domain expert, and this makes learning difficult. With the help of the domain experts, authors can ask more general queries, but avoid the uncertain answers. The contribution of this paper is that the authors demonstrate experimentally that this new method poses fewer queries than is the case with earlier systems of active learning and it can be readily deployed in real-world data mining tasks where obtaining labeled examples is costly.

Two papers in this special section also demonstrate the different applications of $D^3M$ with performance evaluations in the real world of data mining. In the paper, "Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring," Jing He, Yanchun Zhang, Yong Shi, and Guangyan Huang provide a novel domain-driven classification method that takes advantage of multiple criteria and multiple constraint-level programming for intelligent credit scoring. In particular, the domain knowledge, as well as parameters derived from experts' experience, is considered with criteria and constraint functions of linear programming. By measurement and evaluation, feedback is obtained for in-depth modeling. In this step, human interaction is invoked to choose suitable classification parameters, as well as to remodel after monitoring and tracking the process, and also to determine whether the users accept the scoring model. The contribution of this paper is that a novel domain driven data mining approach makes the final credit scoring results simple, meaningful, and easily-controlled by users.

The application addressed in the other paper is a medical one concerning Adverse Drug Reactions (ADRs), a leading cause of hospitalization and death in the world. In "Signaling Potential Adverse Drug Reactions from Administrative Health Databases," Huidong Jin, Jie Chen, Hongxing He, Chris Kelman, Damien McAullay, and Christine M. O'Keefe have studied systematically generated signals of ADRs from administrative health databases. To overcome unexpected and infrequent ADRs, the authors integrated domain intelligence into an ADR knowledge representation: Unexpected Temporal Association Rule (UTAR), with interestingness measures and two successful mining algorithms. Their techniques have great potential to enhance current ADR signaling systems that depend on spontaneous ADR case databases.

The guest editors would like to thank all of the authors who submitted papers for this special section. We also thank all of the referees for their tireless work to deliver quality reviews within deadlines. We believe that this collection of contributions, only possible due to the input of so many people, will play a key role in the future development of Domain-Driven Data Mining research and applications.

Chengqi Zhang
Philip S. Yu
David Bell
*Guest Editors*

**Chengqi Zhang** received the PhD degree from Queensland University in 1991, followed by a Doctor of Science (DSc-Higher Doctorate) from Deakin University in 2002. He has been a research professor in information technology at The University of Technology, Sydney (UTS) since December 2001. He is currently the director of the UTS Research Centre for Quantum Computation and Intelligent Systems. In addition, he is the leader of the data mining program at the Australian Capital Market Cooperative Research Centre. Dr. Zhang's research interests mainly focus on data mining and its applications, especially domain driven data mining, negative association rule mining, and multidatabase mining. He has published more than 200 research papers, including several in first-class international journals, such as *Artificial Intelligence* and IEEE and ACM Transactions. He has delivered 12 keynote/invited speeches at international conferences over the last six years. He has been chairman of the Australian Computer Society National Committee for Artificial Intelligence since November 2005. He is a fellow of the Australian Computer Society (ACS) and a senior member of the IEEE Computer Society. His personal web page can be found at: http://www-staff.it.uts.edu.au/~chengqi/.

**Philip S. Yu** received the BS degree in electrical engineering from National Taiwan University, the MS and PhD degrees in electrical engineering from Stanford University, and the MBA degree from New York University. He is a professor in the Department of Computer Science at the University of Illinois at Chicago and also holds the Wexler Chair in Information Technology. He spent most of his career at the IBM Thomas J. Watson Research Center and was manager of the Software Tools and Techniques Group. His research interests include data mining, Internet applications and technologies, database systems, multimedia systems, parallel and distributed processing, and performance modeling. Dr. Yu has published more than 560 papers in refereed journals and conferences. He holds or has applied for more than 300 US patents. He is a fellow of the ACM and the IEEE. He is an associate editor of the *ACM Transactions on the Internet Technology* and the *ACM Transactions on Knowledge Discovery from Data*. He has received several IBM honors, including two IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, two Research Division Awards, and the 94th plateau of Invention Achievement Awards. He was an IBM Master Inventor.

**David Bell** graduated in 1969 in pure mathematics, and has three subsequent research degrees in a variety of topics: programming language design, database performance, and AI in database systems. He has been a full professor since 1986, currently at Queen's University, Belfast, where he is research director of the Knowledge and Data Engineering research cluster. He has well over 300 publications, including coauthoring *Distributed Databases* (Addison-Wesley) and *Evidence Theory and Its Applications* (North Holland), and has supervised about 35 PhDs to completion. He has been a prime investigator on many national projects and on EU-funded projects (e.g., MAP, ESPRIT, DELTA, COST and AIM) in IT since 1981. He serves/has served on a number of editorial boards, and has chaired / cochaired several program committees, including VLDB '93. His research is in data and knowledge management; the linking of reasoning under uncertainty, machine learning, and other artificial intelligence techniques with more established database work.