Domain-Driven Data Mining: A Practical Methodology

Longbing Cao, University of Technology, Australia

Chengqi Zhang, University of Technology, Australia

ABSTRACT

Extant data mining is based on data-driven methodologies. It either views data mining as an autonomous data-driven, trial-and-error process or only analyzes business issues in an isolated, case-by-case manner. As a result, very often the knowledge discovered generally is not interesting to real business needs. Therefore, this article proposes a practical data mining methodology referred to as domain-driven data mining*, which targets actionable knowledge discovery in a constrained environment for satisfying user preference. The domain-driven data mining consists of a DDID-PD framework that considers key components such as constraint-based context, integrating domain knowledge, human-machine cooperation, in-depth mining, actionability enhancement, and iterative refinement process. We also illustrate some examples in mining actionable correlations in Australian Stock Exchange, which show that domain-driven data mining has potential to improve further the actionability of patterns for practical use by industry and business.*

Keywords: actionable knowledge discovery; constraints; domain-driven data mining; domain knowledge

INTRODUCTION

Extant data mining is presumed as an automated process that produces automatic algorithms and tools without human involvement and the capability to adapt to external environment constraints. As a result, many patterns are mined, but few are workable in real business.

However, actionable knowledge discovery can afford important grounds to business decision makers for performing appropriate actions. In the panel discussions of SIGKDD 2002 and 2003 (Ankerst, 2002, Fayyad et al 2003), it was highlighted by the panelists as one of the grand challenges for extant and future data mining. This situation partly resulted from the scenario that extant data mining is a data-driven trial-and-error process (Ankerst, 2002) in which data mining algorithms extract patterns from converted data via some predefined models based on experts' hypotheses.

Data mining in the real world (e.g., financial data mining and crime pattern mining) (Bagui, 2006) is highly constraint-based (Boulicaut & Jeudy, 2005; Fayyad & Shapiro, 2003). Constraints involve technical, economic, and social aspects in the process of developing and deploying actionable knowledge. Real-world business problems and requirements often are embedded tightly in domain-specific business processes and business rules in charge of expertise (domain constraint). Patterns that are actionable to business often are hidden in large quantities of data with complex data structures, dynamics, and source distribution (data constraint). Often, mined patterns are not actionable to business, even though they are sensible to research. There may be big interestingness conflicts or gaps between academia and business (interestingness constraint). Furthermore, interesting patterns often cannot be deployed to real life, if they are not integrated with business rules, regulations, and processes (deployment constraint). Some other types of constraints include knowledge type constraint, dimension/level constraint, and rule constraint (Han, 1999).

For actionable knowledge discovery from data embedded with the previous constraints, it is essential to slough off the superficial and capture the essential information from the data mining. However, this is a nontrivial task. While many methodologies have been studied, they either view data mining as an automated process or deal with real-world constraints in a case-by-case manner.

Our experience (Cao & Dai, 2003a, 2003b) and lessons learned in data mining in capital markets (Lin & Cao, 2006) show that the involvement of domain knowledge and experts, the consideration of constraints, and the development of in-depth patterns are essential for filtering subtle concerns while capturing incisive issues. Combining these aspects, a sleek data mining methodology can be developed in order to find the distilled core of a problem. It can advise the process of real-world data analysis and preparation, the selection of features, the design and fine-tuning of algorithms, and the evaluation and refinement of mined results in a manner more effective to business. These are our motivations to develop a practical data mining methodology referred to as domain-driven data mining.

Domain-driven data mining consists of a domain-driven in-depth pattern discovery (DDID-PD) framework. The DDID-PD takes I^3D (i.e., interactive, in-depth, iterative, and domain-specific) as real-world KDD bases. $\mathrm{I}^3\mathrm{D}$ means that the discovery of actionable knowledge is an iteratively interactive in-depth pattern discovery process in domain-specific context. $\mathrm{I}^3\mathrm{D}$ is further embodied through (1) mining constraint-based context, (2) incorporating domain knowledge through human-machine-cooperation, (3) mining in-depth patterns, (4) enhancing knowledge actionability, and (5) supporting loop-closed iterative refinement in order to enhance knowledge actionability. Mining constraint-based context requests to effectively extract and transform domain-specific datasets with advice from domain experts and their knowledge.

In the DDID-PD framework, data mining and domain experts complement each other in regard to in-depth granularity through interactive interfaces. The involvement of domain experts and their knowledge can assist in developing highly effective domain-specific data mining techniques and can reduce the complexity of the knowledge-producing process in the real world. In-depth pattern mining discovers more interesting and actionable patterns from a domain-specific perspective. A system following the DDID-PD framework can embed effective supports for domain knowledge and experts' feedback, and refines the life cycle of data mining in an iterative manner.

Taking financial data mining as an example, this article introduces some case studies that deploy the domain-driven data mining methodology. Deep correlations in stock markets are mined through parallel computing to provide measurable benefits for trading. It shows that domain-driven data mining can benefit the actionable knowledge mining in a more effective and efficient manner than data-driven methodology such as CRISP-DM (CRISP).

The remainder of this article is organized as follows. The second section discusses actionable knowledge discovery. The third section presents the DDIP-DM framework. In the fourth section, key components in domain-driven data mining are stated. The fifth section demonstrates case studies on mining actionable correlations in stock markets. We conclude this article and present future work in the sixth section.

ACTIONABLE KNOWLEDGE DISCOVERY

One of the fundamental objectives of KDD is to discover knowledge of main interest to real business needs and user preference. However, this presents a big challenge to extant and future data mining research and applications. Before talking about actionable knowledge discovery, a prerequisite is about what is knowledge actionability. Then, further research can be on developing methodologies and facilities in order to support the discovery of actionable knowledge.

KDD Challenge: Mining Actionable Knowledge

Discovering actionable knowledge has been viewed as the essence of KDD. However, even up to now, it is still one of the great challenges to extant and future KDD, as pointed out by the panel of SIGKDD 2002 and 2003 (Ankerst, 2002) and retrospective literature (Chen & Liu, 2005). This situation partly results from the limitation of extant data mining methodologies, which view KDD as a data-driven, trial-and-error process targeting automated hidden knowledge discovery (Ankerst, 2002; Cao & Zhang, 2006). The methodologies do not take the constrained and dynamic environment of KDD into much consideration, which naturally excludes human and problem domain from the loop. As a result, very often, data mining research mainly aims at developing, demonstrating, and pushing the use of specific algorithms, while it runs off the rails in producing actionable knowledge of main interest to specific user needs.

To revert to the original objectives of KDD, the following three key points recently have been highlighted: comprehensive constraints around the problem (Boulicaut & Jeudy, 2005), domain knowledge (Pohle, Yoon, Henschen, Park, & Makki, 1999), and human role (Ankerst, 2002; Cao & Dai, 2003a; Han, 1999) in the process and environment of real-world KDD. A proper consideration of these aspects in the KDD process has been reported to make KDD promising in digging out actionable knowledge satisfying real life dynamics and requests, even though this is very tough issue. This pushes us to think of what is knowledge actionablility and how to support actionable knowledge discovery.

We further study a practical methodology called domain-driven data mining for actionable knowledge discovery (Cao & Zhang, 2006). On top of the data-driven framework, domain-driven data mining aims to develop proper methodologies and techniques for integrating domain knowledge, human role and interaction, as well as actionability measures into the KDD process in order to discover actionable knowledge in the constrained environment. This research is very important for developing the next-generation data mining methodology and infrastructure (Ankerst, 2002; Cao & Zhang, 2006). It can assist in a paradigm shift from data-driven hidden pattern mining to domain-driven actionable knowledge discovery, and provides supports for KDD to be translated to real business situations, as widely expected.

Knowledge Actionability

Often, mined patterns are nonactionable to real needs due to the interestingness gaps between academia and business (Gur & Wallace, 1997). Therefore, measuring actionability of knowledge is essential in order to recognize interesting links that permit users to react to them to better service business objectives. The measurement of knowledge actionability should be from perspectives of both objective and subjective.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items, *DB* be a database that consists of a set of transactions, and *x* be an itemset in *DB*. Let *P* be an interesting pattern discovered in *DB* through utilizing a model *M*. The following concepts are developed for the DDID-PD framework.

Definition 1. Technical Interestingness. The technical interestingness *tech_int*() of a rule or a pattern is highly dependent on certain technical measures of interest specified for a data mining method. Technical interestingness is measured further in terms of technical objective measures *tech_obj*() and technical subjective measures *tech_sub*().

Definition 2. Technical Objective Interestingness. Technical objective measures *tech_obj*() capture the complexities of a link pattern and its statistical significance. It could be a set of criteria. For instance, the following logic formula indicates that an association rule *P* is technically interesting if it satisfies *min_support* and *min_confidence*.

$\forall x \in I, \exists P : x.min\_support(P) \wedge x.min\_confidence(P) \rightarrow x.tech\_obj(P)$

Definition 3. Technical Subjective Interestingness. On the other hand, technical subjective measures *tech_subj*(), also focusing and based on technical means, recognize to what extent a pattern is of interest to a particular user's needs. For instance, probability-based belief (Padmanabhan & Tuzhilin, 1998) is developed for measuring the expectedness of a link pattern.

Definition 4. Business Interestingness. The business interestingness *biz_int*() of an itemset or a pattern is determined from domain-oriented social, economic, user preference and/or psychoanalytic aspects. Similar to technical interestingness, business interestingness also is represented by a collection of criteria from both objective *biz_obj*() and subjective *biz_subj*() perspectives.

Definition 5. Business Objective Interestingness. The business objective interestingness *biz_obj*() measures to what extent the findings satisfy the concerns from business needs and user preference based on objective criteria. For instance, in stock trading pattern mining, profit and roi (return on investment) often is used for judging the business potential of a trading pattern objectively. If the profit and roi (return on investment) of a stock price predictor *P* are satisfied, then *P* is interesting to trading.

$\forall x \in I, \exists P : x.profit(P) \wedge x.roi(P) \rightarrow x.biz\_obj(P)$

Definition 6. Business Subjective Interestingness. *Biz_subj*() measures business and user concerns from subjective perspectives such as psychoanalytic factors. For instance, in stock trading pattern mining, a kind of psycho-index 90% may be used to indicate that a trader thinks it is very promising for real trading.

A successful discovery of an actionable knowledge is a collaborative work between miners and users, which satisfies both academia-oriented technical interestingness measures *tech_obj*() and *tech_subj*() and domain-specific business interestingness *biz_obj*() and *biz_subj*().

Definition 7. Actionability of a Pattern. Given a pattern *P*, its actionable capability *act*() is described as to what degree it can satisfy both the technical and business interestingness.

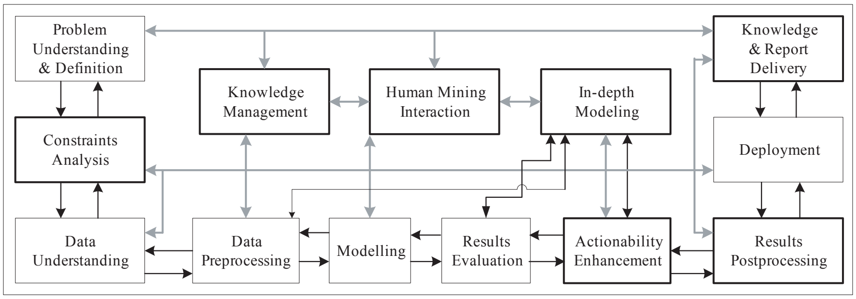
$\forall x \in I, \exists P : act(P) = f(tech\_obj(P) \wedge tech\_subj(P) \wedge biz\_obj(P) \wedge biz\_subj(P))$

If a pattern is discovered automatically by a data mining model while it only satisfies technical interestingness request, it usually is called an (technically) interesting pattern. It is presented as:

$\forall x \in I, \exists P : x.tech\_int(P) \rightarrow x.act(P)$

In a special case, if both technical and business interestingness, or a hybrid interestingness measure integrating both aspects, are satisfied, it is called an actionable pattern. It is not only interesting to data miners but generally interesting to decision makers.

Figure 1. DDID-PD process model



$$\forall x \in I, \exists P : x.tech\_int(P) \wedge x.biz\_int(P) \rightarrow x.act(P)$$

Therefore, the work of actionable knowledge discovery must focus on knowledge findings that can satisfy not only technical interestingness but also business measures.

DDID-PD FRAMEWORK

The existing data mining methodology (e.g., CRISP) generally supports autonomous pattern discovery from data. The DDID-PD, on the other hand, highlights a process that discovers in-depth patterns from constraint-based context with the involvement of domain experts/knowledge. Its objective is to maximally accommodate both naïve users as well as experienced analysts and satisfy business goals. The patterns discovered are expected to be actionable to solve domain-specific problems and can be taken as grounds for performing effective actions. In order to make domain-driven data mining effective, user guides and intelligent human-machine interaction interfaces are essential through incorporating both human qualitative intelligence and machine quantitative intelligence. In addition, appropriate mechanisms are required to deal with multiform constraints and domain knowledge. This section outlines key ideas and relevant research issues of DDID-PD.

DDID-PD Process Model

The main functional components of the DDID-PD are shown in Figure 1, in which we highlight those processes specific to DDID-PD in thick boxes. The life cycle of DDID-PD is as follows, but be aware that the sequence is not rigid; some phases may be bypassed or moved back and forth in a real problem.

Every step of the DDID-PD process may involve domain knowledge and the assistance of domain experts.

P1. Problem understanding
P2. Constraints analysis
P3. Analytical objective definition, feature construction
P4. Data preprocessing
P5. Method selection and modeling or
P5'. In-depth modeling
P6. Initial generic results analysis and evaluation
P7. It is quite possible that each phase from P1 may be iteratively reviewed through analyzing constraints and interaction with domain experts in a back-and-forth manner or
P7': In-depth mining on the initial generic results where applicable
P8. Actionability measurement and enhancement
P9. Back and forth between P7 and P8
P10. Results post-processing
P11. Reviewing phases from P1 may be required

P12. Deployment
P13. Knowledge delivery and report synthesis for smart decision making

The DDID-PD process highlights the following highly correlated ideas that are critical for the success of a data mining process in the real world. They are as follows:

1. **Constraint-Based Context.** Actionable pattern discovery is based on a deep understanding of the constrained environment surrounding the domain problem, data, and its analysis objectives.
2. **Integrating Domain Knowledge.** Real-world data applications inevitably involve domain and background knowledge, which is very significant for actionable knowledge discovery.
3. **Cooperation Between Human and Data Mining System.** The integration of human role and the interaction and cooperation between domain experts and mining systems in the whole process are important for effective mining execution.
4. **In-Depth Mining.** Another round of mining on the first-round results may be necessary for searching patterns really interesting to business.
5. **Enhancing Knowledge Actionability.** Based on the knowledge actionability measures, the actionable capability of findings needs to be further enhanced from modeling and evaluation perspectives.
6. **Loop-Closed Iterative Refinement.** Patterns actionable for smart business decision making in most cases would be discovered through loop-closed iterative refinement.
7. **Interactive and Parallel Mining Supports.** It is necessary and helpful to develop business-friendly system supports for human-mining interaction and parallel mining for complex data mining applications.

The following section outlines each of them respectively.

KEY COMPONENTS SUPPORTING DOMAIN-DRIVEN DATA MINING

In domain-driven data mining, the following seven key components are advocated. They have potential for making KDD different from the existing data-driven data mining if they are appropriately considered and supported from technical, procedural, and business perspectives.

Constraint-Based Context

In human society, everyone is constrained either by social regulations or by personal situations. Similarly, actionable knowledge only can be discovered in a constraint-based context such as environmental reality, expectations, and constraints in the mining process. Specifically, in the first section, we list several types of constraints that play significant roles in a process, effectively discovering knowledge actionable to business. In practice, many other aspects, such as data stream and the scalability and efficiency of algorithms, may be enumerated. They consist of domain-specific, functional, nonfunctional, and environmental constraints. These ubiquitous constraints form a constraint-based context for actionable knowledge discovery. All of the previous constraints to varying degrees must be considered in relevant phases of real-world data mining. In this case, it is even called constraint-based data mining (Boulicaut & Jeudy, 2005; Han, 1999).

Some major aspects of domain constraints include the domain and characteristics of a problem, domain terminology, specific business process, policies and regulations, particular user profiling, and favorite deliverables. Potential matters to satisfy or react on domain constraints could consist of building domain model, domain metadata, semantics, and ontologies (Cao, Zhang, & Liu, 2005); supporting human involvement, human-machine interaction, qualitative and quantitative hypotheses, and conditions; merging with business processes and enterprise information infrastructure; fitting regulatory measures; conducting user profile analysis and modeling; and so forth. Relevant hot research areas include interactive mining, guided mining, knowledge and human involvement, and so forth.

Constraints on particular data may be embodied in terms of aspects such as very large volume, ill structure, multimedia, diversity, high dimensions, high frequency and density, distribution and privacy, and so forth. Data constraints seriously affect the development of and performance requirements on mining algorithms and systems, and constitute some grand challenges to data mining. As a result, some popular researches on data constraints-oriented issues are emerging, such as stream data mining, link mining, multi-relational mining, structure-based mining, privacy mining, multimedia mining, and temporal mining.

What makes this rule, pattern, and finding more interesting than the other? In the real world, simply emphasizing technical interestingness such as objective statistical measures of validity and surprise is not adequate. Social and economic interestingness (we refer to Business Interestingness) such as user preferences and domain knowledge should be considered in assessing whether a pattern is actionable or not. Business interestingness would be instantiated into specific social and economic measures in terms of the problem domain. For instance, profit, return and roi usually are used by traders to judge whether a trading rule is interesting enough or not.

Furthermore, the delivery of an interesting pattern must be integrated with the domain environment such as business rules, process, information flow, presentation, and so forth. In addition, many other realistic issues must be considered. For instance, a software infrastructure may be established to support the full life cycle of data mining; the infrastructure needs to integrate with the existing enterprise information systems and workflow; parallel KDD may be involved with parallel supports on multiple sources, parallel I/O, parallel algorithms, and memory storage; visualization, privacy, and security should receive much deserved attention; and false alarming should be minimized.

In summary, actionable knowledge discovery won't be a trivial task. It should be put into a constraint-based context. On the other hand, tricks not only may include how to find a right pattern with a right algorithm in a right manner, but they also may involve a suitable process-centric support with a suitable deliverable to business.

Integrating Domain Knowledge

It is accepted (Pohle et al., 1999) gradually that domain knowledge can play significant roles in real-world data mining. For instance, in trading pattern mining, traders often take "beating market" as a personal preference to judge an identified rule's actionability. In this case, a stock mining system needs to embed the formulas calculating market return and rule return, and set an interface in order for traders to specify a favorite threshold and comparison relationship between the two returns in the evaluation process. Therefore, the key is to take advantage of domain knowledge in the KDD process.

The integration of domain knowledge is subject to how it can be represented and filled in to the knowledge discovery process. Ontology-based domain knowledge representation, transformation, and mapping between business and data mining systems is one of the proper approaches (Cao et al., 2005) to model domain knowledge. Further work is to develop agent-based cooperation mechanisms (Cao et al., 2004; Zhang, Zhang, & Cao, 2005) to support ontology-represented domain knowledge in the process.

Domain knowledge in the business field often takes forms of precise knowledge, concepts, beliefs, relations, or vague preference and bias. Ontology-based specifications build a business ontological domain to represent domain knowledge in terms of ontological items and semantic relationships. For instance, in the previous example, return-related items include return, market return, rule return, and so forth. There is *class_of* relationship between return and market return, while market return is associated with rule return in some form of user-specified logic connectors, say beating market if rule return is larger than (>) market return by a threshold ϕ. We can develop ontological representations to manage these items and relationships.

Further, business ontological items are mapped to a data mining system's internal ontologies. So, we build a mining ontological domain for a KDD system collecting standard domain-specific ontologies and discovered knowledge. To match items and relationships between two domains and to reduce and aggregate synonymous concepts and relationships in each domain, ontological rules, logical connectors, and cardinality constraints are studied in order to support the ontological transformation from one domain to another and the semantic aggregations of semantic relationships and ontological items' intra- or interdomains. For instance, the following rule transforms ontological items from the business domain to the mining domain. Given input item A from users, if it is associated with B by *is_a* relationship, then the output is B from the mining domain: $\forall$ (A AND B), $\exists$ B ::= *is_a*(A, B) $\Rightarrow$ B, the resulting output is B. For rough and vague knowledge, we can fuzzify and map them to precise terms and relationships. For the aggregation of fuzzy ontologies, fuzzy aggregation and defuzzification mechanisms can be developed in order to sort out proper output ontologies.

Cooperation Between Human and Mining Systems

The real requirements for discovering actionable knowledge in constraint-based context determine that real-world data mining is more likely to be human involved than automated. Human involvement is embodied through the cooperation among humans (including users and business analysts, mainly domain experts) and data mining systems. This is achieved through the complementation between human qualitative intelligence, such as domain knowledge and field supervision, and mining quantitative intelligence like computational capability. Therefore, real-world data mining likely presents as a human-machine-cooperated interactive knowledge discovery process.

The role of humans can be embodied in the full period of data mining from business and data understanding, problem definition, data integration and sampling, feature selection, hypothesis proposal, business modeling, and the evaluation, refinement, and interpretation of algorithms and resulting outcomes. For instance, experience, metaknowledge, and imaginary thinking of domain experts can guide or assist with the selection of features and models, adding business factors into the modeling, creating high-quality hypotheses, designing interestingness measures by injecting business concerns, and quickly evaluating mining results. This assistance largely can improve the effectiveness and efficiency of mining actionable knowledge.

Humans often serve on the feature selection and result evaluation. Humans may play roles in a specific stage or during the full stages of data mining. Humans can be an essential constituent of or the center of a data mining system. The complexity of discovering actionable knowledge in constraint-based context determines to what extent a human must be involved. As a result, the human-mining cooperation could be, to varying degrees, human-centered or guided mining (Ankerst, 2002; Fayyad & Shapiro, 2003), or human-supported or assisted mining, and so forth.

In order to support human involvement, human mining interaction or, in a sense, presented as interactive mining (Aggarwal, 2002; Ankerst, 2002) is absolutely necessary. Interaction often takes explicit forms; for instance, setting up direct interaction interfaces to fine tune parameters. Interaction interfaces may take various forms as well, such as visual interfaces; virtual reality techniques; multi-modal, mobile agents, and so forth. On the other hand, it also could go through implicit mechanisms; for example, accessing a knowledge base or communicating with a user assistant agent. Interaction communication may be message-based, model-based, or event-based. Interaction quality relies on performance such as user-friendliness, flexibility, run-time capability, representability, and even understandability.

Mining In-Depth Patterns

The situation that many mined patterns are interesting more to data miners than to businesspersons has hindered the deployment and adoption of data mining in real applications. Therefore, it is essential to evaluate the actionability of a pattern and to further discover actionable patterns; namely, $\forall P: x.tech\_int(P) \wedge x.biz\_int(P) \rightarrow x.act(P)$, to support smarter and more effective decision making. This leads to *in-depth pattern mining*.

Mining in-depth patterns should consider how to improve both technical (*tech_int()*) and business interestingness (*biz_int()*) in the previous constraint-based context. Technically, it could be through enhancing or generating more effective interestingness measures (Omiecinski, 2003); for instance, a series of research has been done on designing right interestingness measures for association rule mining (Tan, Kumar, & Srivastava, 2002). It also could be through developing alternative models for discovering deeper patterns. Some other solutions include further mining actionable patterns on the discovered pattern set. Additionally, techniques can be developed in order to deeply understand, analyze, select, and refine the target data set in order to find in-depth patterns.

More attention should be paid to business requirements, objectives, domain knowledge, and qualitative intelligence of domain experts for their impact on mining deep patterns. This can be through selecting and adding business features, involving domain knowledge into modeling, supporting interaction with users, tuning parameters and data set by domain experts, optimizing models and parameters, adding factors into technical interestingness measures or building business measures, improving result evaluation mechanisms through embedding domain knowledge and human involvement, and so forth.

Enhancing Knowledge Actionability

Patterns that are interesting to data miners may not lead necessarily to business benefits, if deployed. For instance, a large number of association rules often is found, while most of them might not be workable in business situations. These rules are generic patterns or technically interesting rules. Further actionability enhancement is necessary for generating actionable patterns of use to business.

The measurement of actionable patterns is to follow the actionablilty of a pattern. Both technical and business interestingness measures must be satisfied from both objective and subjective perspectives. For those generic patterns identified based on technical measures, business interestingness needs to be checked and emphasized so that the business requirements and user preference can be put into proper consideration.

Actionable patterns in most cases can be created through rule reduction, model refinement, or parameter tuning by optimizing generic patterns. In this case, actionable patterns are a revised optimal version of generic patterns that capture deeper characteristics and understanding of the business and are also called in-depth or optimized patterns. Of course, such patterns also can be directly discovered from a data set with sufficient consideration of business constraints. For instance, the section "Mining Actionable Trading Rules" discusses mining actionable trading rules from a great number of generic rules.

Loop-Closed Iterative Refinement

Actionable knowledge discovery in a constraint-based context is likely to be a closed rather than an open process. It encloses iterative feedback to varying stages such as sampling, hypothesis, feature selection, modeling, evaluation, and interpretation in a human-involved manner. On the other hand, real-world mining process is highly iterative, because the evaluation and refinement of features, models, and outcomes cannot be completed once but, rather, is based on iterative feedback and interaction before reaching the final stage of knowledge and decision-support report delivery.

The previous key points indicate that real-world data mining cannot be dealt with just an algorithm; rather, it is really necessary to build a proper data mining infrastructure in order to discover actionable knowledge from constraint-based scenarios in a loop-closed iterative manner. To this end, agent-based data mining infrastructure (Klusch et al., 2003; Zhang et al., 2005) presents good facilities, since it provides good supports for autonomous problem-solving, user modeling, and user agent interaction.

Interactive and Parallel Mining Supports

To support domain-driven data mining, it is significant to develop interactive mining supports for human-mining interaction and to evaluate the findings. On the other hand, parallel mining supports often are necessary and can greatly upgrade the real-world data mining performance.

For interactive mining supports, intelligent agents and service-oriented computing are some good technologies. They can support flexible, business-friendly, and user-oriented human-mining interaction through building facilities for user modeling; user knowledge acquisition; domain knowledge modeling; personalized user services and recommendation; run-time supports; and mediation and management of user roles, interaction, security, and cooperation.

Based on our experience in building agent service-based stock trading and mining system F-Trade (Cao et al., 2004; F-TRADE), an agent service-based actionable discovery system can be built for domain-driven data mining. User agent, knowledge management agent, ontology services (Cao et al., 2005), and run-time interfaces can be built to support interaction with users, take users' requests, and manage information from users in terms of ontologies. Ontology-represented domain knowledge and user preferences then are mapped to mining domain for mining purposes. Domain experts can help to train, supervise, and evaluate the outcomes.

Parallel (Domingos, 2003; Taniar & Rahayu, 2002) and scalable (Manlatty & Zaki, 2000) KDD supports involve parallel computing and management supports to deal with multiple sources, parallel I/O, parallel algorithms, and memory storage. For instance, in order to tackle cross-organization transactions, we can design efficient parallel KDD computing and system supports in order to wrap data mining algorithms. This can be through developing parallel genetic algorithms and proper processor-cache memory techniques. Multiple master-client, process-based genetic algorithms and caching techniques can be tested on different CPU and memory configurations in order to find good parallel computing strategies.

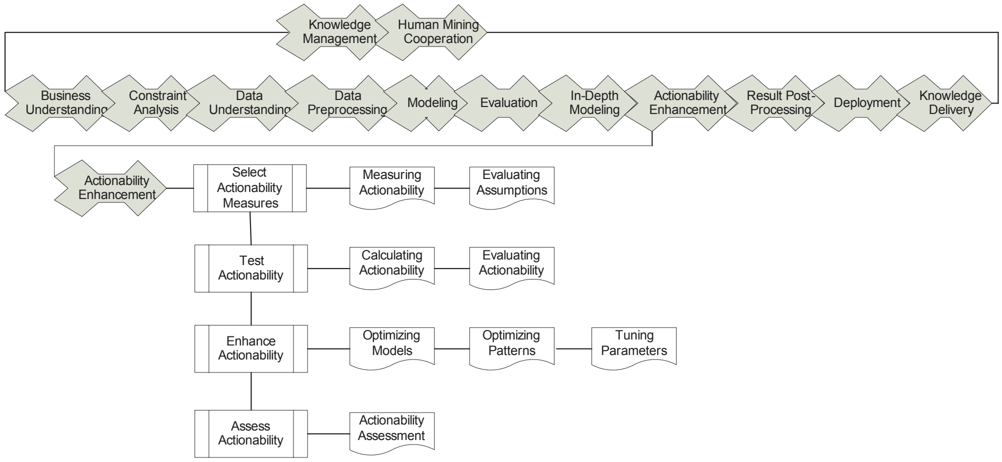
The facilities for interactive and parallel mining supports largely can improve the performance of real-world data mining in aspects such as human-mining interaction and cooperation, user modeling, domain knowledge capturing, reducing computation complexity, and so forth. They are some essential parts of next-generation KDD infrastructure.

Reference Model and Questionnaire

Reference models such as those in CRISP-DM are very helpful for guiding and managing the knowledge discovery process. It is recommended that those reference models be respected in domain-oriented, real-world data mining. However, actions and entities for domain-driven data mining, such as considering constraints and integrating domain knowledge, should be paid special attention in the corresponding models and procedures. On the other hand, new reference models are essential for supporting components such as in-depth modeling and actionablility enhancement. For instance, Figure 2 illustrates the reference model for actionability enhancement.

In the field of developing real-world data mining applications, questionnaires are very

Figure 2. Actionability enhancement



helpful for capturing business requirements, constraints, requests from organization and management, risk and contingency plans, expected representation of the deliverables, and so forth. It is recommended to design questionnaires for every procedure in the domain-driven actionable knowledge discovery process.

Reports for every procedure must be prepared and recorded in the knowledge management base for organizing well the knowledge and the process of domain-driven data mining applications.

DOMAIN-DRIVEN MINING APPLICATIONS

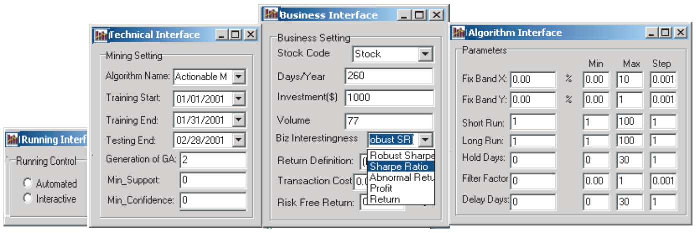
In this section, we illustrate some of our work in financial data mining (Lin & Cao, 2006) by utilizing domain-driven data mining methodologies. We only highlight some of those key components such as domain knowledge, in-depth rule mining, business interestingness, and parallel mining for pattern pruning in financial trading evidence discovery.

Financial data mining (Kovalerchuk & Vityaev, 2000) is of high interest, since it may benefit trading decision and market surveillance, but it also may be challenging, because financial markets are greatly complex. Taking ASX as an instance, there are more than 1,000 shares listed in this small market. In the Data Mining Program (DMP) of Australian Capital Markets Cooperative Research Center (CMCRC), we deploy the domain-driven data mining methodology to actionable trading evidence discovery, such as mining correlations between stocks, actionable trading rules, and correlations between trading rules and stocks. The following sections illustrate some results of the previous work in ASX data.

Mining Correlated Stocks

In real trading, traders often trade multiple stocks in order to manage risk. Data mining may extract evidences about what stocks are correlated with others. A common hypothesis is that stocks from the same or similar sectors or belonging to a shared production chain to some extent may be correlated. We have developed a set of correlation metrics in order to analyze the relations between stocks in the ASX. The following outlines the basic idea of mining correlated stocks by considering relevant market factors.

Figure 3. Interfaces supporting human-mining system interaction



Running Interf
Running Control
○ Automated
○ Interactive

Technical Interface
Mining Setting
Algorithm Name: Actionable M
Training Start: 01/01/2001
Training End: 01/31/2001
Testing End: 02/28/2001
Generation of GA: 2
Min_Support: 0
Min_Confidence: 0

Business Interface
Business Setting
Stock Code: Stock
Days/Year: 260
Investment($): 1000
Volume: 77
Biz Interestingness: obust SR
Return Definition: (
Transaction Cost: 0.0
Risk Free Return: 0
Robust Sharpe
Sharpe Ratio
Abnormal Retu
Profit
Return

Algorithm Interface
Parameters

			Min	Max	Step
Fix Band X:	0.00	%	0.00	10	0.001
Fix Band Y:	0.00	%	0.00	1	0.001
Short Run:	1		1	100	1
Long Run:	1		1	100	1
Hold Days:	0		0	30	1
Filter Factor	0		0.00	1	0.001
Delay Days:	0		0	30	1

Algorithm: Mining Correlated Stocks

C1. Calculating the coefficient ρ of two stocks considering market impact

C2. Determining the scope of ρ interesting to trading through cooperation with traders by considering market aspects, such as market sectors, volatility, liquidity, and index

C3. Evaluation by designing and simulating strategies to trade the correlated stocks

C4. Recommending correlated stocks

In the ASX, we targeted 32 stocks with quality data from January 1997 to June 2002. Thirteen of those stocks were found to be highly correlated. Of all the 78 pairs of combinations, nine pairs were found to be actionable to trading with expectable profits. For instance, we found that stock A (representing some stock) is highly correlated with B. The return on trading the pair A-B was 40.51% on average on historical data from January 1, 1997, to June 19, 2002, without considering the market impact.

In mining correlated stocks benefiting trading, we found the following interesting points: (1) Correlated stocks interesting to trading cannot be determined just by coefficient, but rather, market aspects such as sectors, volatility, liquidity, and index should be considered, as well. (2) Interestingly, all correlated stocks mined in the ASX come from different sectors. This finding means that correlated stocks are not necessary from the same industry, as presumed by financial researchers. (3) The return on trading a correlated pair is affected highly by the liquidity and volatility of a stock.

Mining Actionable Trading Rules

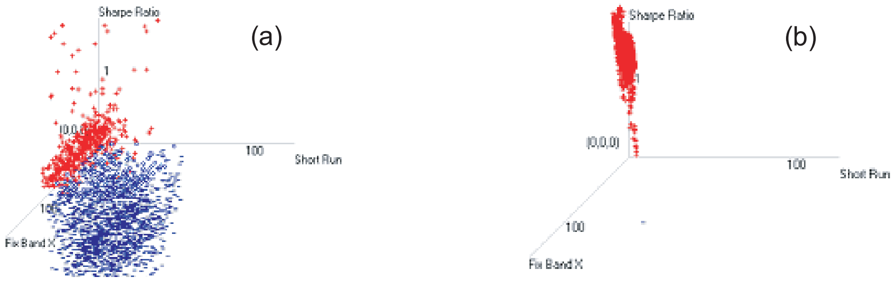
A trading rule actually indicates a possible investment pattern in stock markets. For instance, the trading rule MA (*sr*, *lr*, δ) indicates a correlated trading pattern between features *short-run moving average* (*sr*) and *long-run moving average* (*lr*). The trading strategy is defined as follows (where δ is a fixed difference band between *sr* and *lr*).

IF *sr* *(1-δ) >= *lr* THEN *Buy*
IF *sr* *(1+δ) <= *lr* THEN *Sell*

In market trading, the previous pattern MA actually can be instantiated into millions of individual generic rules such as MA(2, 50, 0.01) and MA(10, 50, 0.01). However, traders do not know which rule is actionable for a specific investment scenario. Therefore, mining actionable trading rules emerge as a worthwhile activity.

In order to involve domain knowledge in finding actionable rules, we built human mining interaction interfaces. Figure 3 demonstrates some interfaces in which users can trigger the process in terms of automated execution or interactive mode with involvement of users. In interactive mode, technical analysts can advise the previous process as well as refine technical factors for mining setting and algorithm parameter tuning. Business analysts can supervise the construction of features, fine tune the parameters, and set evaluation criteria for business concerns. For instance, measure *sharpe_ratio* is used for evaluating the business actionability of an identified rule. Additionally,

Figure 4. Improved business interestingness by in-depth rules: (a) sharpe ratio with generic MA rules and (b) sharpe ratio with actionable MA rules



the system supports ad-hoc execution, meaning that users can tune the parameters or change interestingness measures to check the results at run time.

$$sharpe\_ratio = (r_P - r_R) / \sigma_P$$

where r_P is the expected portfolio return, r_R is risk-free rate, and σ_P is portfolio-standard deviation. Higher *sharpe_ratio* means more return with lower risk.

We found a collection of actionable rules using our actionable trading rule mining algorithms. For instance, in ASX data, MA(4, 19, 0.033) is a very interesting rule using training data from January 1, 2000, to December 31, 2000, and testing set between January 1, 2001, and December 31, 2001. The number of trading signals generated by this rule is much bigger than other possible rules with good *sharpe_ratio*. Figure 4(b) shows that its *sharpe_ratio* has a greatly improved positive scope compared with (a) the generic results. This demonstrates that the in-depth pattern mining with the involvement of domain knowledge can improve the actionability of trading rules.

Mining Rule-Stock Correlations

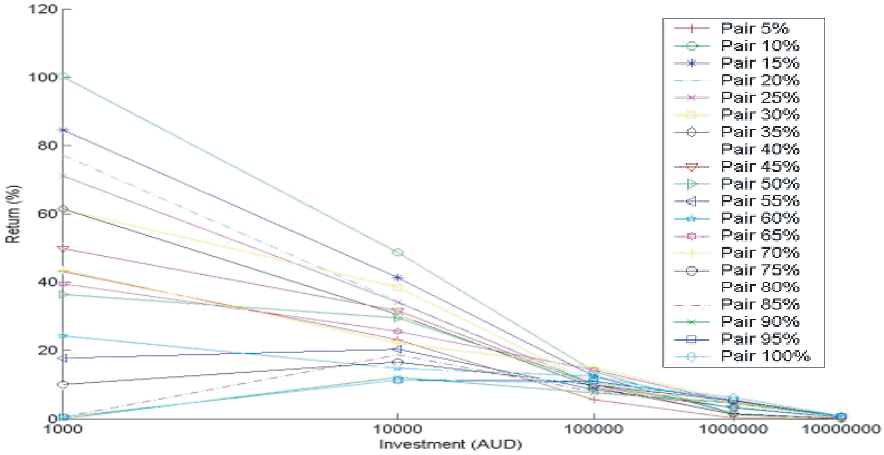
In market, some trading rules are tested to be effective to trade a class of stocks, while other rules are more suitable for other stocks. Using data mining, we may evidence that whether there exist correlations between trading rules and stocks. If we do discover some actionable correlations, then it would be helpful for trading. Based on this hypothesis, we developed algorithms to find the correlations between trading rules and stocks in stock market data. The basic ideas of the rule-stock correlation mining algorithms are as follows.

Algorithm: Mining Correlated Trading Rule-Stocks Pairs

C1. Mining actionable rules for an individual stock
C2. Mining highly correlated rule-stock pairs by high dimension reduction
C3. Evaluating and refining the rule-stock pairs by considering traders' concerns
C4. Recommending actionable rule-stock pairs

In discovering actionable rule-stock pairs, traders were invited to give suggestions on designing features, interestingness measure and parameter optimization. They also helped us to design mechanisms for evaluating and refining rule-stock pairs. Taking the ASX as an instance, three types of trading rules (MA, Filter Rule, and Channel Breakout) (Ryan, Allan, & Halbert, 2005) and 26 ASX stocks were chosen for the experiments. For instance, the intraday training data was from January 1, 2001, to January 31, 2001, and the testing set was

Figure 5. Return on investment with actionable rule-stock pairs



from February 1, 2001, to February 28, 2001. Five investment plans were conducted on the previous rules and stocks. In organizing pairs, we ranked them based on return and generated 5% pair, 10% pair, and so forth from the whole pair set. The 5% pair means that return for trading these pairs is the top 5% in the whole pair set. Figure 5 illustrates returns for different investment plans on different pairs. These graphs are interesting to traders, allowing them to make smart trading decisions using these mined rule-stock pairs.

Parallel Computing

Mining actionable correlations in a scenario with hundreds of stocks (e.g., more than 1,000) and millions of trading rules on stock data with hundreds of thousands of intraday stock transactions (e.g., more than 700,000 per day) is very time-consuming. Parallel computing is essential for acceptable response time. Taking the mining actionable trading rules as an example, we designed different parallel algorithms A_i in order to test their performance on ASX stock C (representing a stock in ASX) using intraday data in 2001.

Alg_1. Loops through all possible combinations of MA (*sr*, *lr*, δ).

Alg_2. Parallelizes Alg_1 by partitioning the search calculations into four processing units.

Alg_3. Parallelizes Alg_1 by partitioning the search calculations into eight processing units.

Alg_4. Parallelizes Alg_1 by splitting processes into master and slave subprocesses on four processing units.

Alg_5. Parallelizes Alg_1 by splitting processes into master and slave subprocesses on eight processing units.

We tested the previous algorithms on a Linux box with eight CPUs (Intel(R) Xeon(TM) MP CPU 2.00GHz) and 4GB memory. The running time for each algorithm is shown in Table 1. The results indicate that parallel computing and efficient implementations can extremely accelerate the computation of data mining. However, in our case, eight CPUs make little difference from four CPUs. This is probably due to the overhead from system and managing master and slave subprocesses.

Table 1. Running time for mining actionable MA

Algorithms	Running Time (seconds)
Alg_1	860
Alg_2	26
Alg_3	22
Alg_4	13
Alg_5	11

CONCLUSIONS AND FUTURE WORK

Real-world data mining applications have proposed urgent requests for discovering actionable knowledge of main interest to real-user and business needs. Actionable knowledge discovery is significant and also very challenging. It is nominated as one of great challenges of KDD in the next 10 years. The research on this issue has potential to change the existing situation in which a great number of rules are mined while few of them are interesting to business, and to promote the wide deployment of data mining into business.

This article has developed a new data mining methodology referred to as Domain-Driven Data Mining. It provides a systematic overview of the issues in discovering actionable knowledge and advocates the methodology of mining actionable knowledge in constraint-based context through human mining system cooperation in a loop-closed iterative refinement manner. It is useful for promoting the paradigm shift from data-driven hidden pattern mining to domain-driven actionable knowledge discovery. Further, progress in studying domain-driven data mining methodologies and applications can help the deployment shift from standard or artificial data set-based testing to real data and business environment-based backtesting and development.

On top of data-driven data mining, domain-driven data mining includes almost all phases of the well-known industrial data mining methodology CRISP-DM. However, it also has enclosed some big differences from the data-driven methodologies, such as CRISP-DM. For instance:

- Some new essential components, such as in-depth modeling, the involvement of domain experts and knowledge, and knowledge actionability measurement and enhancement are taken into the life cycle of KDD for consideration.
- In the domain-driven methodology, the phases of CRISP-DM highlighted by thick boxes in Figure 1 are enhanced by dynamic cooperation with domain experts and the consideration of constraints and domain knowledge.
- Knowledge actionability is highlighted in the discovery process. Both technical and business interestingness must be concerned in order to satisfy needs and especially business requests.

These differences actually play key roles in improving the existing knowledge discovery in a more effective way.

In the deployment of the domain-driven data mining methodology, we have demonstrated some of our research results in mining actionable correlations in Australian stock markets. The experiments show that domain-driven data mining has potential for improving the actionable knowledge mining. Our further work is on developing detailed mining process specifications and interfaces for easily deploying domain-driven data mining methodology into real-world mining.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council (ARC) Discovery Projects (DP0449535 and DP0667060), UTS Chancellor and ECRG funds, National Science Foundation of China (60496327), and Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01). We appreciate CMCRC and SIRCA for providing data services. Thanks also go to Dr. Lin Li and Jiarui Ni for implementation supports.

REFERENCES

Aggarwal, C. (2002). Towards effective and interpretable data mining by visual interaction. *ACM SIGKDD Explorations Newsletter, 3*(2), 11–22.

Ankerst, M. (2002). Report on the SIGKDD-2002 panel the perfect data mining tool: Interactive or automated? *ACM SIGKDD Explorations Newsletter, 4*(2), 110–111.

Bagui, S. (2006). An approach to mining crime patterns. *International Journal of Data Warehousing and Mining, 2*(1), 50–80.

Boulicaut, J-F., & Jeudy, B. (2005). Constraint-based data mining. In O. Maimon, & L. Rokach (Eds.), *The data mining and knowledge discovery handbook* (pp. 399–416). New York: Springer.

Cao, L., & Dai, R. (2003a). Human-computer cooperated intelligent information system based on multi-agents. *ACTA Automatica Sinica, 29*(1), 86–94.

Cao, L., & Dai, R. (2003b). Agent-oriented metasynthetic engineering for decision making. *International Journal of Information Technology and Decision Making, 2*(2), 197–215.

Cao, L. et al. (2004). Agent services-based infrastructure for online assessment of trading strategies. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Intelligent Agent Technology* (pp. 345–349). IEEE Press.

Cao, L., & Zhang, C. (2006). Domain-driven actionable knowledge discovery in the real world. In *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2006, LNAI 3918* (pp. 821–830). Singapore: Springer.

Cao, L., Zhang, C., & Liu, J. (2005). *Ontology-based integration of business intelligence* (Technical Report). Sydney, Australia: University of Technology, E-Intelligence Group.

Chen, S.Y., & Liu, X. (2005). Data mining from 1994 to 2004: an application-orientated review. *International Journal of Business Intelligence and Data Mining, 1*(1), 4–21.

CMCRC (Captial Markets Cooperative Research Centre). (n.d.). Retrieved from http://www.cmcrc.com

CRISP. (n.d.). Retrieved from http://www.crisp-dm.org

Domingos, P. (2003). Prospects and challenges for multi-relational data mining. *SIGKDD Explorations, 5*(1), 80–83.

Fayyad, U., Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel—Data mining: The next 10 years. *ACM SIGKDD Explorations Newsletter, 5*(2), 191–196.

F-TRADE. (n.d.). Retrieved from http://www.f-trade.info

Gur Ali, O.F., & Wallace, W.A. (1997). Bridging the gap between business objectives and parameters of data mining algorithms. *Decision Support Systems, 21*, 3–15.

Han, J. (1999). *Towards human-centered, constraint-based, multi-dimensional data mining* [Speech]. Minneapolis, MN: University of Minnesota.

Klusch, M. et al. (2003). The role of agents in distributed data mining: Issues and benefits. In *Proceedings of the 2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT 2003)* (pp. 211–217). IEEE Computer Society Press.

Kovalerchuk, B., & Vityaev, E. (2000). *Data mining in finance: Advances in relational and hybrid methods*. Massachusetts: Kluwer Academic Publishers.

Lin, L., & Cao, L. (2006). *Mining in-depth patterns in stock market* (Technical Report). Sydney, Australia: University of Technology, E-Intelligence Group.

Manlatty, M., & Zaki, M. (2000). Systems support for scalable data mining. *SIGKDD Explorations, 2*(2), 56–65.

Omiecinski, E. (2003). Alternative interest measures for mining associations. *IEEE Transactions on Knowledge and Data Engineering, 15*, 57–69.

Padmanabhan, B., & Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 94–100). Menlo Park, CA: AAAI Press.

Pohle, C. (n.d.). *Integrating and updating domain knowledge with data mining*. Retrieved from http://citeseer.ist.psu.edu/668556.html

Ryan, S., Allan, T., & Halbert, W. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Financial, 54*(5), 1647–1692.

Tan, P., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 32–41). ACM Press.

Taniar, D., & Rahayu, J.W. (2002). Parallel data mining. In H.A. Abbass, R. Sarker, & C. Newton (Eds.), *Data mining: A heuristic approach* (pp. 261–289). Hershey, PA: Idea Group Publishing.

Yoon, S., Henschen, L., Park, E., & Makki, S. (1999). Using domain knowledge in knowledge discovery. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*. ACM Press.

Zhang, C., Zhang, Z., & Cao, L. (2005). Agents and data mining: Mutual enhancement by integration. In *Proceedings of the International Workshop on Autonomous Intelligent Systems: Agents and Data Mining, LNAI 3505* (pp. 50–61). Berlin: Springer.

Dr. Longbing Cao, IEEE Senior Member, has been heavily involved in research, commerce and leadership related to KDD. He has served as PC member in more than 10 international conferences, and as chief technical officer, CI, and team or program leader in Australia and China. He has published over 50 refereed papers in data mining and multi-agent systems, among other areas of research. He has proven knowledge, experience and leadership in approximately 10 large KDD-related research and commercial projects. He has delivered data mining services to areas such as capital markets, telecom industries and governmental services.

Professor Chengqi Zhang, IEEE Senior Member, is currently a research professor in Faculty of Information Technology at University of Technology, Sydney. His areas of research are data mining and multi-agent systems. He has published more than 200 refereed papers, edited nine books, and published three monographs. He served as an associate editor or a member of the editorial board for five international journals. He has served as general chair, PC chair, or organizing chair for four international conferences and a member of program committees for many international or national conferences. He has attracted more than $1 million research grants.