International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 21, No. 3 (2007) 1–16

THE EVOLUTION OF KDD: TOWARDS DOMAIN-DRIVEN DATA MINING*

LONGBING CAO† and CHENGQI ZHANG‡

Faculty of Information Technology
University of Technology, Sydney, Australia
†*lbcao@it.uts.edu.au*
‡*chengqi@it.uts.edu.au*

Traditionally, data mining is an autonomous data-driven trial-and-error process. Its typical task is to let data tell a story disclosing hidden information, in which domain intelligence may not be necessary in targeting the demonstration of an algorithm. Often knowledge discovered is not generally interesting to business needs. Comparably, real-world applications rely on knowledge for taking effective actions. In retrospect of the evolution of KDD, this paper briefly introduces *domain-driven data mining* to complement traditional KDD. *Domain intelligence* is highlighted towards actionable knowledge discovery, which involves aspects such as domain knowledge, people, environment and evaluation. We illustrate it through mining activity patterns in social security data.

Keywords: Data mining; knowledge actionability; domain-driven data mining.

1. Introduction

In the last decade, data mining, or KDD (knowledge discovery in database),[13] has become an active research and development area in information technology fields. In particular, data mining is gaining rapid development in various aspects such as the data mined, the knowledge discovered, the techniques developed, and the applications involved. Table 1 illustrates such key research and development progress in KDD.

A typical feature of traditional data mining is that KDD is presumed as an automated process. It targets the production of automatic algorithms and tools. As a result, algorithms and tools developed have no capability to adapt to external environment constraints. Millions of patterns and algorithms are published in academia but unfortunately very few of them have been transferred into real business.

Many researchers and developers have realized the limitation of traditional data mining methodologies, and the gap between business and academic attention.

*This work was sponsored by Australian Research Council Discovery Grant (DP0667060, DP0773412), China Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01), and UTS internal grants.

Table 1. Data mining development.

Dimension	Key Research Progress
Data mined	— Relational, data warehouse, transactional, object-relational, active, spatial, time-series, heterogeneous, legacy, WWW
	— Stream, spatiotemporal, multimedia, ontology, event, activity, links, graph, text, etc.
Knowledge discovered	— Characters, associations, classes, clusters, discrimination, trend, deviation, outliers, etc.
	— Multiple and integrated functions, mining at multiple levels, mining exceptions, etc.
Techniques developed	— Database-oriented, association and frequent pattern analysis, multidimensional and OLAP analysis methods, classification, cluster analysis, outlier detection, machine learning, statistics, visualization, etc.
	— Scalable data mining, stream data mining, spatiotemporal data and multimedia data mining, biological data mining, text and Web mining, privacy-preserving data mining, event mining, link mining, ontology mining, etc.
Application involved	— engineering, retail market, telecommunication, banking, fraud detection, intrusion detection, stock market, etc.
	— Specific task-oriented mining
	— Biological, social network analysis, intelligence and security, etc.
	— Enterprise data mining, cross-organization mining, etc.

The research on challenges of KDD and innovative and workable KDD methodologies and techniques has actually become a significant and productive direction of KDD. In the panel discussions of SIGKDD 2002 and 2003,[2,9] a couple of grand challenges for extant and future data mining were identified. Among them, for instance, actionable knowledge discovery is one of key focuses, because it can not only afford important grounds to business decision makers for performing appropriate actions, but also deliver expected outcomes to business. However, it is not a trivial task to extract actionable knowledge utilizing traditional KDD methodologies. This situation partly results from the scenario that extant data mining is a data-driven trial-and-error process,[2] where data mining algorithms extract patterns from converted data through predefined models based on experts' hypothesis.

To bridge the gap between business and academia, it is important to understand the difference of objectives and goals of data mining in research and in real world. Real-world data mining presents extra constraints and expectation on mined results, for instance, financial data mining and crime pattern mining is highly constraint-based.[3,9] The difference gets involved in key aspects such as the concerned problems, context mined KDD, interested patterns, the processes of mining, cared interests, and infrastructure supporting data mining.

To handle the above difference, real-world experience[4,5] and lessons learned in data mining in capital markets[16] show the significance of domain intelligence.

Domain intelligence consists of the involvement of domain knowledge[25] and experts, the consideration of constraints, and the development of in-depth patterns, which are essential for filtering subtle concerns while capturing incisive issues. Combining these together, a sleek data mining methodology is necessary to find the distilled core of a problem. They form the grounds of *domain-driven data mining*.

This paper provides a view of rethinking traditional data mining towards real-world actionable knowledge discovery. The remainder of this paper is organized as follows. Section 2 discusses the evolution of KDD. Section 3 presents major criteria for measuring the actionability of knowledge. In Sec. 4, key components in domain-driven data mining are stated. Section 5 briefly states domain-driven data mining framework. A case study is demonstrated in Sec. 6. We conclude this research and present future work in Sec. 7.

2. KDD: Data Driven versus Domain Driven

One of the fundamental objectives of KDD is to discover knowledge of main interest to real business needs and user preference. This forms a big challenge to extant and future data mining research and applications. To better understand this conflict, let us review traditional data-driven data mining methodologies and research, and the expectation of read world KDD.

2.1. *Extant data mining: Data-driven interesting pattern discovery*

Conceptually, there is no problem with traditional data mining, which views data mining as a process of data-driven interesting pattern discovery. After all, data mining targets useful information hidden in data. However, attention there has been simply or mainly paid to data itself. This may be evidenced by the research scope, methodologies, and research interest of traditional data mining. We may generate a picture of traditional data mining by summarizing its major characteristics from the following aspects: (i) object mined: data is the object being mined, which is expected to tell the whole story of a concern; (ii) aims of data mining are to develop innovative approaches in this period. As a result of this motivation and trend, almost all high-level papers talk about new approaches; (iii) datasets mined are abstract or refined from real problems or data. Mining is not directly conducted on raw data from business; (iv) correspondingly, the objective of data mining is to develop or update and demonstrate new algorithms on a very nice data set; (v) models and methods in data mining systems are usually predefined. It is the data mining researcher rather than a user that can deploy an algorithm; (vi) the process of data mining is packed as automated, in which a user is not necessary and actually he/she cannot do much in the mining procedure; (vii) the evaluation of mined results is basically based on technical metrics, if a threshold presumed by data mining researchers is higher, then the algorithm is promising; (viii) among (vii) the accuracy of an algorithm is taken as one of key criteria of quality judgment.

In a summary, traditional KDD is a data-driven trial-and-error process targeting automated hidden knowledge discovery.[2,7] The goal of traditional data mining is to let data create/verify research innovation, demonstrate and push the use of novel algorithms discovering knowledge of interest to researchers.

2.2. *Real world KDD: Domain-driven actionable knowledge discovery*

In the real world, discovering knowledge actionable in solving problems concerned 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[2,9] and retrospective literature. This situation partly results from the limitation of traditional data mining methodologies, which do not take into much consideration the constrained and dynamic environment of KDD. They naturally exclude human and problem domain in the loop of data mining. As a result, very often data mining research mainly aims at developing, demonstrating and pushing the use of specific algorithms. As a result, it runs off the rails in producing actionable knowledge of main interest to specific user needs.

In the wave of rethinking original objectives of KDD, the following key points have recently been highlighted: comprehensive constraints around a problem,[3] domain knowledge and human role[2,4,12] in KDD process and environment. A proper consideration of these aspects in the KDD process has been reported to make KDD promising to dig out actionable knowledge satisfying real life dynamics and requests even though this is a very tough issue. This pushes us to think of what knowledge actionablility is, and how to support actionable knowledge discovery.

Aiming to complement the shortcoming of traditional data mining, in particular, satisfying the real user needs in enterprise data mining, we study a practical methodology, called *domain-driven data mining*.[7] The basic theory of domain-driven data mining is as follows. On top of the data-driven framework, it aims to develop proper methodologies and techniques for integrating domain knowledge, human role and interaction, as well as actionability measures into KDD process. It targets to discover actionable knowledge in a practical constrained environment. This research is very important for developing the next-generation data mining methodology and infrastructure.[2,7] 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 the real business situations as widely expected.

In contrast with the traditional data mining, we also list the content of domain-driven data mining research and development. Most importantly, in domain-driven data mining, it is data and domain intelligence (including domain knowledge and domain experts) that work together to tell a hidden story in business, which discovers actionable knowledge to satisfy real user needs. It is the user who say "yes" or "no" to mined results. Table 2 compares major aspects under research of traditional data-driven and domain-driven data mining.

1st Reading

May 7, 2007 16:0 WSPC/115-IJPRAI SPI-J068 00561

Table 2. Data-driven versus domain-driven data mining.

Aspects	Traditional Data-Driven	Domain-Driven
Object mined	Data tells the story	Data and domain (business rules, factors etc.) tell the story
Aim	Developing innovative approaches	Generating business impacts
Objective	Algorithms are the focus	Systems are the target
Dataset	Mining abstract and refined data set	Mining constrained real life data
Extendibility	Predefined models and methods	Ad-hoc, running-time and personalized model customization
Process	Data mining is an automated process	Human is in the circle of data mining process
Evaluation	Evaluation based on technical metrics	Business say "yes" or "no"
Accuracy	Accurate and solid theoretical computation	Data mining is a kind of artwork
Goal	Let data create/verify research innovation; Demonstrate and push the use of novel algorithms discovering knowledge of interest to research	Let data and domain knowledge tell hidden story in business; discovering actionable knowledge to satisfy real user needs

3. What Makes KDD of Interest to Business

In traditional data mining, often mined patterns are nonactionable to real needs due to gaps of interests between academia and business.[11] Therefore, it is critical to get a clear answer to the problem "what makes KDD of interest to business".[20] Answers to it may be quite varying. Basically, traditional data mining focuses on developing and refining *technical objective* measures. A typical example is those metrics developed for associations.[22] Recently, *subjective* metrics are also paid attention by researchers. On the other hand, domain-driven data mining verifies and validates the usability of a pattern based not only on *technical* measures but also on *business* concerns. A more likely scenario is to integrate technical concerns with business ones, and generate an integrative measurement system to justify the quality of mined results. To this end, the concept of knowledge actionability is essential for recognizing interesting links permitting users to react to them to better service business objectives. The measurement of *knowledge actionability* should be from both *objective* and *subjective* perspectives. Table 3 summarizes the measurement of interest of data-driven versus domain-driven data mining.

Table 3. Measurement of interest of data-driven versus domain-driven data mining.

Interest		Traditional Data-Driven	Domain-Driven
Technical	Objective	Technical objective $tech\_obj()$	Technical objective $tech\_obj()$
	Subjective	Technical subjective $tech\_subj()$	Technical subjective $tech\_subj()$
Business	Objective	—	Business objective $biz\_obj()$
	Subjective	—	Business subjective $biz\_subj()$
Integrative		—	Actionability $act()$

3.1. *Technical significance versus business expectation*

The development of actionability is a progressive process in data mining. In the framework of traditional data mining, the so-called actionability is mainly embodied in terms of technical significance. *Technical interesting* $tech\_int()$ measures whether a pattern is of interest or not in terms of specific statistical significance corresponding to a particular data mining method. There are two steps in technical interest evolution. The original focus basically was on *technical objective interest* $tech\_obj()$,[10,14] which aims to capture the complexities of pattern structure and statistical significance. For instance, *coefficient* is developed for measuring objective interest of correlated stocks. Recent work appreciated *technical subjective* measures $tech\_sub()$,[17,19,21] which also recognize to what extent a pattern is of interest to a particular user. For example, probability-based belief is used to describe user confidence of unexpected rules.[19]

Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of items, DB be a database consisting of transactions, x is an itemset in DB. Let P be interesting evidence discovered in DB through a modeling method M. For the above two procedures, we have the follows.

Phase 1: $\forall x \in X$, $\exists P$: $x.tech\_obj(P) \longrightarrow x.act(P)$

Phase 2: $\forall x \in X$, $\exists P$: $x.tech\_obj(P) \wedge x.tech\_subj(P) \longrightarrow x.act(P)$

Gradually, data miners realize that the actionability of a discovered pattern must be assessed by and satisfies domain user needs. To achieve business expectations, *business interestingness* $biz\_int()$ measures to what degree a pattern is of interest to a business person from social, economic, personal and psychoanalytic factors. Similar to $tech\_int()$, recently *business objective interest* $biz\_obj()$ is recognized by some researchers, say profit mining[24] and domain-driven data mining,[7] involving $biz\_int()$. At this stage, we get Phase 3 as:

Phase 3: $\forall x \in X$, $\exists P$: $x.tech\_obj(P) \wedge x.tech\_subj(P) \wedge x.biz\_obj(P) \longrightarrow x.act(P)$

Moreover, *business subjective interest* $biz\_sub()$ also plays essential roles in assessing $biz\_int()$. This leads to a comprehensive cognition of actionability as indicated by Phase 4 advocated in domain-driven data mining.

Phase 4: $\forall x \in X$, $\exists e$: $x.tech\_obj(P) \wedge x.tech\_subj(P) \wedge x.biz\_obj(P) \wedge x.biz\_subj(P) \longrightarrow x.act(P)$

3.2. *Knowledge actionability*

Based on the above assessment, knowledge actionability should highlight both academic and business concerns.[7] *Actionability* recognizes technical significance of an extracted pattern that also permits users to specifically react to it to better service their business objectives. Since the satisfaction of technical interest is the antecedent of actionability, we view actionable knowledge as what satisfies not only technical interestingness $tech\_int()$ but also user-specified business interest $biz\_int()$. We have the following definition.

Definition 1. (Knowledge Actionability) Given a mined pattern, its actionable capability *act(P)* is described as the degree of its satisfaction with both technical and business interests.

$$\forall x \in X, \exists P : x.tech\_int(P) \wedge biz\_int(P) \longrightarrow act(P) \tag{1}$$

Further, it is instantiated in terms of objective and subjective factors from both technical and business sides.

$$\begin{aligned}\forall x \in X, \exists P : x.tech\_obj(P) \wedge x.tech\_subj(P) \wedge x.biz\_obj(P)\\ \wedge biz\_subj(P) \longrightarrow act(P)\end{aligned} \tag{2}$$

In this case, there are two sets of interest measures needed to be calculated when a pattern is extracted. For instance, we say a mined association trading rule is (technically) interesting because it satisfies requests on *support* and *confidence*. Moreover, if it also beats the expectation of user-specified market index return *IR* then it is a generally actionable rule.

In the real-world mining, business interests $biz\_int()$ may differ or conflict technical significance $tech\_int()$. The relationship between them may present as one of four scenarios as listed in Table 4.

Clearly, actionable knowledge mining targets patterns confirming the relationship $tech\_int() \Leftrightarrow biz\_int()$. However, it is a kind of artwork to tune thresholds and balance significance and difference between $tech\_int()$ and $biz\_int()$. Quite often a pattern with high $tech\_int()$ creates bad $biz\_int()$. Contrarily, it is not a rare case that a pattern with low $tech\_int()$ generates good $biz\_int()$. In this case, it is domain users who can better tune thresholds and difference. Besides the above-discussed work on developing useful technical and business interest measures, there are some other things to do to reach and enhance knowledge actionability such as efforts on selecting actionability measures, testing actionability, enhancing actionability and assessing actionability in domain-driven data mining process.[7]

4. Towards Domain Driven Data Mining

Data mining research and development is boosted by challenges from the real world. For instance, some typical recent progress made in data mining includes stream

Table 4. Relationship between technical significance and business expectation.

Relationship Type	Explanation
$tech\_int() \Leftarrow biz\_int()$	The pattern e does not satisfy business expectation but technical significance
$tech\_int() \Rightarrow biz\_int()$	The pattern e does not satisfy technical significance but business expectation
$tech\_int() \Leftrightarrow biz\_int()$	The pattern e satisfies business expectation as well as technical significance
$tech\_int() \not\Leftrightarrow biz\_int()$	The pattern e satisfies neither business expectation nor technical significance

data mining handling stream data, link mining studying linkage across entities. Challenges and prospects coming from the real world force us to rethink some key points in data mining. This includes problem understanding and definition, KDD context, patterns mined, mining process, interest system, and infrastructure supports. The outcome of this retrospection and rethinking is a paradigm shift from traditional data-driven-focused research towards domain-driven-oriented research and development. The domain-driven data mining has potential for making KDD available for satisfying real user needs rather than demonstrating algorithms if relevant points can be appropriately considered and supported from technical, procedural and business perspectives.

4.1. *Problem: Domain-free versus domain-specific*

In traditional data mining, researchers pay a large amount of time in constructing research problems, which in real-world data mining comes from real challenges. As a typical phenomenon, even though a problem may come from a real scenario, it is always abstracted and pruned into a very general and brilliant research issue to fill in innovation and significance requirements. Such research issue is usually domain-free, which means it does not necessarily involve specific domain intelligence. Undoubtedly, this is important for developing the science of KDD.

On the other hand, in real-world scenarios, challenges always come from specific domain problems. Therefore, objectives and goals of applying KDD are basically problem-solving to satisfy real user needs. Problem-solving and satisfying real user needs present strongly usable requirements. Requirements mainly come from a specific domain involving concrete functional and nonfunctional concerns. The analysis and modeling of these requirements request domain intelligence, in particular domain background knowledge and involvement of domain experts. Therefore, real-world data mining is more likely domain-specific. However, domain-specific data mining is not necessarily specific domain-problem oriented. Here *domain* can refer to either a big industrial sector, for instance, telecom or banking, or a categorical business such as customer relationship management.

Domain intelligence can play significant roles in real-world data mining. Domain knowledge in business field often takes forms of precise knowledge, concepts, beliefs, relations, or vague preference and bias. For instance, in cross-market mining, traders often take "beating market" as a personal preference to judge an identified rule's actionability. The key to taking advantage of domain knowledge in the KDD process is knowledge and intelligence integration, which involves how it can be represented and filled into the knowledge discovery process. Ontology-based domain knowledge representation, transformation and mapping between business and data mining system is a proper approach to model domain knowledge. Ontology-based specifications build a *business ontological domain* to represent domain knowledge in terms of ontological items and semantic relationships. Ontological representation[6] can be developed to manage the above items and relationships.

Through ontology-based representation and transformation, business terms are mapped to data mining system's internal ontologies. We build an internal *data mining ontological domain* for KDD system collecting standard domain-specific terms and discovered knowledge. To match items and relationships between two domains and reduce and aggregate synonymous concepts and relationships in each domain, ontological rules, logical connectors and cardinality constraints are studied to support ontological transformation from one domain to another, and semantic aggregations of semantic relationships and ontological items intra or inter domains.

4.2. *KDD context: Unconstrained versus constrained*

Law, business rule and regulation are common forms of constraints in human society. Similarly, actionable knowledge discovery can only be well conducted in a constrained rather than unconstrained context. Constraints involve technical, economic and social aspects in the process of developing and deploying actionable knowledge. For instance, constraints can be something involving aspects such as environmental reality and expectations on data format, knowledge representation, and outcome delivery in the mining process. Other aspects of domain constraints include domain and characteristics of a problem, domain terminology, specific business process, policies and regulations, particular user profiling and favorite deliverables. In particular, we highlight following types of constraints — *domain constraint*, *data constraint*, *interest constraint* and *deployment constraint*.

The real-world business problems and requirements are often tightly embedded in domain-specific business process and business rules (*domain constraint*). Potential matters to satisfy or react on domain constraints may consist of building domain model, domain metadata, semantics and ontologies,[6] 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, etc. Relevant hot research areas include interactive mining, guided mining, and knowledge and human involvement etc.

Patterns that are actionable to business are often hidden in large quantities of data with complex data structures, dynamics and source distribution (*data constraint*). 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, etc. Data constraints seriously affect the development and performance requirements of 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, multirelational mining, structure-based mining, privacy mining, multimedia mining and temporal mining.

Often mined patterns are not actionable to business even though they are sensible to research. There may be huge conflicts of interest or gaps between academia and business (*interest constraint*). What makes this rule, pattern and finding more interesting than the other? In the real world, simply emphasizing technical interest such as objective statistical measures of validity and surprise is not adequate. Social and economic interests (we refer to Business Interests) such as user preferences and domain knowledge should be considered in assessing whether a pattern is actionable or not. Business interests may be instantiated into specific social and economic measures in terms of a problem domain. For instance, *profit*, *return* and *roi* are usually used by traders to judge whether a trading rule is interesting enough or not.

Furthermore, often interesting patterns cannot be deployed to real life if they are not integrated with business rules and processes (*deployment constraint*). The delivery of an interesting pattern must be integrated with the domain environment such as business rules, process, information flow, presentation, etc. In addition, many other realistic issues must be considered. For instance, a software infrastructure may be established to support the full lifecycle of data mining; the infrastructure needs to integrate with the existing enterprise information systems and workflow; parallel KDD[23] may be involved with parallel supports on multiple sources, parallel I/O, parallel algorithms, memory storage; visualization, privacy and security should receive much-deserved attention; false alarms should be minimized.

Some other types of constraints include knowledge type constraint, dimension/level constraint and rule constraint.[12] Several types of constraints play significant roles in effectively discovering knowledge actionable to business world. In practice, many other aspects such as data stream and 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 the above constraints must, to varying degrees, be considered in relevant phases of real-world data mining. In this case, it is even called *constraint-based data mining*.[3,12]

4.3. *Pattern: Generic versus actionable patterns*

Many mined patterns are more useful to data miners than to business persons. Generally interesting patterns are useful because they satisfy technical interest measurement. These rules are *generic patterns* or technically interest rules.

However, they are not necessarily useful for solving business problems. To improve this situation, we advocate in-depth pattern mining which aims to develop patterns actionable in business world. It targets the discovery of actionable patterns to support smart and effective decision-making, namely a pattern P must satisfy

$$\forall P : x.tech\_int(P) \wedge x.biz\_int(P) \longrightarrow x.act(P). \tag{3}$$

Therefore, in-depth patterns can be delivered through improving either technical interests *tech_int*() or business interests *biz_int*(). As discussed in Sec. 3 on pattern interests, both technical and business interest measures must be satisfied from both objective and subjective perspectives.

Technically, it could be through enhancing or generating more effective interest measures.[18] For instance, a series of research have been done on designing right interest measures for association rule mining.[22] It may also be through developing alternative models for discovering deeper patterns. Some other solutions include further mining actionable patterns on a discovered pattern set. Additionally, techniques can be developed to deeply understand, analyze, select and refine the target data set in order to find in-depth patterns. 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, which capture deeper characteristics and understanding of the business. Of course, such patterns can also be directly discovered from data set with sufficient consideration of business constraints.

On the other hand, for those generic patterns identified based on technical measures, their business interest needs to be checked so that business requirements and user preference can be put into proper consideration. Domain intelligence, including business requirements, objectives, domain knowledge and qualitative intelligence of domain experts, can play roles in enhancing pattern actionability. This can be achieved through selecting and adding business features, involving domain knowledge, supporting interaction with users, tuning parameters and data set by domain experts, optimizing models and parameters, adding factors into technical interest measures or building business measures, improving result evaluation mechanism through embedding domain knowledge and human involvement.

4.4. *Infrastructure: Automated versus human-mining-cooperated*

Traditional data mining is an automated trial and error process. Deliverables are presumed as automated predefined algorithms and tools. It is arguable that such automated methodology has both strengths and weaknesses. The good side is to make user life easy. However, it meets with challenges such as a lack of capability in involving domain intelligence and adapting to dynamic situations in the business world. In particular, automated data mining has trouble in handling enterprise data mining applications.

Actionable knowledge discovery in constrained context determine that real-world data mining is more likely to be human involved rather than automated. Human involvement is embodied through the cooperation between human (including users and business analysts, mainly domain experts) and data mining system. 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 is likely to be present as a human-machine-cooperated interactive knowledge discovery process.

Human role 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 learning to evaluation, refinement and interpretation of algorithms and resulting outcomes. For instance, experience, metaknowledge and imaginary thinking of domain experts can guide or assist with selection of features and models, adding business factors into modeling, creating high quality hypotheses, designing interest measures by injecting business concerns, and quickly evaluating mining results. This assistance can largely improve the effectiveness and efficiency of mining actionable knowledge.

Usually, humans serve on feature selection and result evaluation. Humans can play roles in a specific stage or full stages of data mining. Humans can be an essential constituent or the centre of data mining system. The complexity of discovering actionable knowledge in constraint-based context decides to what extent and how humans must be involved. As a result, human-mining cooperation presents to varying degrees, human-centered, guided mining,[29] or human-assisted mining.

To support human involvement, human mining interaction, or perhaps presented as interactive mining,[1,2] 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 technique, multimodal, agents,[15] etc. On the other hand, it could also 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, presentable capability and understandability.

5. Domain-Driven KDD Framework

We have presented a domain-driven data mining framework.[7] Domain-driven data mining consists of the following key components (i) problem understanding and the definition is domain-specific and must involve domain intelligence, (ii) data mining is in a constraint-based context, (iii) pattern discovery targets mining in-depth patterns, (iv) data mining presented as a loop-closed iterative refinement process, (v) the mined results must be actionable in business, and (vi) building a human-machine-cooperated infrastructure supporting domain-driven data mining. In domain-driven framework, data mining and domain experts complement each other with regards 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 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 this framework can embed effective supports for domain knowledge and experts' feedback, and refine the lifecycle of data mining in an iterative manner.

6. Case Study

Here we briefly illustrate the development of actionable activity patterns in social security data[8] using domain-driven data mining. Taking frequent activity sequence mining as an instance, we identify those i-itemset ($i = 2, 3, 4, \ldots$) frequent activity sequences likely associated with the occurrence of government customer debt using sequential association mining. Due to the imbalance of class and item distribution of debt-related activities, we split activities into two classes with domain supervision: debt-related activity set and nondebt related activity set. To handle such unbalanced data, we develop both technical and business metrics for measuring the actionability of a pattern. The following technical metrics are defined: *global support*, *local support*, *class difference rate*, *relative risk ratio*.

Definition 2. The global support of a pattern $\{P \longrightarrow \$\}$ in activity set A is defined as $Supp_A(P, \$) = |P, A|/|A|$.

If $Supp_A(P, \$)$ is larger than a given threshold, then P is a frequent activity sequence in A leading to debt. $Supp_A(P, \$)$ reflects the global statistical significance of the rule $\{P \longrightarrow \$\}$ in activity set A.

Definition 3. The local support (L_SUPP) of a rule $\{P \longrightarrow \$\}$ in target activity set D is defined as $Supp_A(P, \$) = |P, D|/|D|$. On the other hand, the local support of rule $\{P \longrightarrow \overline{\$}\}$ in activity set $A - D$ (i.e. nondebt activity set) is defined as $Supp_{A-D}(P, \overline{\$}) = |P, A - D|/|A - D|$. The class difference rate $Cdr(P, |^D_{A-D})$ of P in two independent classes D and $A - D$ is defined as:

$$Cdr(P, |^D_{A-D}) = Supp_D(P, \$)/Supp_{A-D}(P, \overline{\$}). \tag{4}$$

If $Cdr(P, |^D_{A-D})$ is larger than a given threshold, then P far more frequently leads to debt than nondebt. This measure indicates the difference between targeted class and untargeted class. An obvious difference between them is expected for positive frequent impact-targeted activity patterns.

Definition 4. Given local support ($SUPP$) $Supp_D(P, \$)$ and $SUPP_{A-D}(P, \overline{\$})$, the relative risk ratio $Rrr(P, |^{\$}_{\overline{\$}})$ of P leading to target activity classes D and nontarget class $A - D$ is defined as:

$$Rrr(P, |^{\$}_{\overline{\$}}) = Prob(\$|P)/Prob(\overline{\$}|P) = Prob(P, \$)/Prob(P, \overline{\$}) \tag{5}$$

$$Rrr(P, |^{\$}_{\overline{\$}}) = Supp_A(P, \$)/Supp_A(P, \overline{\$}) \tag{6}$$

If $Rrr(P, |^{\$}_{\overline{\$}})$ is larger than a given threshold, then P far more frequently leads to debt than results in nondebt. This indicates statistical difference of a sequence P leading to debt or nondebt in a global manner. An obvious difference between them is expected to distinguish frequent impact-targeted activity patterns. In addition, if the statistical significance of P leading to $\$$ and $\overline{\$}$ are compared in terms of local

Table 5. Technical interest metrics in activity sequence mining in social security area.

PATTERN	LSUPP	SUPP	CONF	LIFT	ZSCORE	$Cdr(P, \|_{A-D}^{D})$	$Rrr(P, \|_{\$}^{\$})$
$A, E \longrightarrow DET$	0.0186	0.0157	0.845	1.69	3.73		

classes, then relative risk ratio $Rrr(P, |_{\$}^{\$})$ indicates the difference of a pattern's significance between targeted class and untargeted class as defined in Definition 4.

A number of sequential activity patterns are mined based on the above and traditional measures such as left side support (LSUPP), confidence (CONF), lift (LIFT) and z score (ZSCORE). For instance, the following Table 5 illustrates one sequential activity pattern ($A, E \longrightarrow DET$) likely associated with debt in balanced mix data (where A and E are activity labels).

We then prune this pattern set by developing business interest metrics, for instance, the following specify the impact of a mined activity sequence on averaged debt amount and debt duration: *pattern average debt amount*, and *pattern average debt duration*.

Definition 5. The total debt amount $d\_amt()$ is the sum of all individual debt amounts $d\_amt_i (i = 1, \ldots, f)$ in f itemsets holding the pattern ACB. Then we get pattern average debt amount for the pattern ACB:

$$\overline{d\_amt}() = \sum_{1}^{f} d\_amt()_i / f \tag{7}$$

Definition 6. Debt duration $d\_dur()$ for pattern ACB is the average duration of all individual debt durations in f itemsets holding ACB. Debt duration $d\_dur()$ of an activity is the number of days a debt keeps valid, $d\_dur() = d.end\_date$ - $d.start\_date$ *+1*, where $d.end\_date$ is the day a debt is completed, $d.start\_date$ is the day a debt is activated. Pattern average debt duration $\overline{d\_dur}()$ is defined as:

$$\overline{d\_dur}() = \sum_{1}^{f} d\_dur()_i / f \tag{8}$$

For instance, the following lists technical and business interest measures of activity sequence rule "$L, O \longrightarrow DET$" for Australian social security benefit recipients. If the activity "O" follows "L" in customer contacts, then the customer is likely to be in government customer debt. The technical interest tells the statistical significance of this rule, while business interest shows governmental officers how important this rule leads to debt cost to the Government.

- Technical interest:
  - support = 0.01251
  - confidence = 0.60935
  - lift = 1.2187

- Business interest:
  - $\overline{d\_amt}() = 29{,}526$, the averaged debt amount in cents of those debt-related activity sequences supporting the rule
  - $\overline{d\_dur}() = 15.5$, the averaged debt duration in days of those debt-related activity sequences supporting the rule

7. Conclusions and Future Work

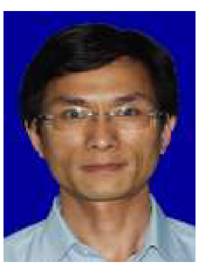
The retrospection of traditional data mining has disclosed the significance of developing KDD methodologies and supports targeting actionable knowledge discovery. Domain-driven data mining provides complementary supports and ideas on traditional data-driven data mining. It adequately utilizes domain intelligence including domain expertise, knowledge, constraints, environment, human cooperation for deep and actionable pattern mining satisfying business expectation fitting in business rules and processes.

Domain-driven data mining has been used in telecom data mining, financial data mining and government service mining. They have shown that it has a potential to strengthen traditional KDD where a great number of rules are mined while few of them are interesting to business, and promote a wide deployment of data mining into business. Our further work is performed on qualitative analysis of the impact of domain intelligence on KDD, as well as the representation and integration of domain knowledge into KDD systems.

References

1. C. Aggarwal, Towards effective and interpretable data mining by visual interaction, *ACM SIGKDD Explor. Newslett.* **3**(2) (2002) 11–22.
2. M. Ankerst, Report on the SIGKDD-2002 panel the perfect data mining tool: interactive or automated? *ACM SIGKDD Explor. Newslett.* **4**(2) (2002) 110–111.
3. J. F. Boulicaut and Jeudy, B. Constraint-based data mining, *The Data Mining and Knowledge Discovery Handbook* (Springer, 2005), pp. 399–416.
4. L. Cao and R. Dai, Human-computer cooperated intelligent information system based on multi-agents, *ACTA AUTOMATICA SINICA* **29**(1) (2003) 86–94.
5. L. Cao and R. Dai, Agent-oriented metasynthetic engineering for decision making, *Int. J. Inform. Technol. Dec. Mak.* **2**(2) (2003) 197–215.
6. L. Cao *et al.*, Ontology-based integration of business intelligence, *Web Intell. Age. Syst.: an Int. J.* **4**(4) (2006) (to appear).
7. L. Cao *et al.*, Domain-driven data mining: a practical methodology, *Int. J. Data Warehousing and Mining* **2**(4) (2006) 49–65.
8. L. Cao *et al.*, Mining impact-targeted activity patterns in unbalanced data, Technical Report, University of Technology Sydney, 2006.
9. U. Fayyad, G. Shapiro and R. Uthurusamy, Summary from the KDD-03 panel — Data mining: the next 10 years, *ACM SIGKDD Explor. Newslett.* **5**(2) (2003) 191–196.
10. A. A. Freitas, On objective measures of rule surprisingness, *PKDD98*, 1998, pp. 1–9.
11. O. F. Gur Ali and W. A. Wallace, Bridging the gap between business objectives and parameters of data mining algorithms, *Decision Support Syst.* **21** (1997) 3–15.

12. J. Han, Towards human-centered, constraint-based, multi-dimensional data mining, An invited talk at Univ. Minnesota, Minneapolis, Minnesota, 1999.
13. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd edition (Morgan Kaufmann, 2006).
14. R. J. Hilderman and H. J. Hamilton, Applying objective interestingness measures in data mining systems, *PKDD00*, 2000, pp. 432–439.
15. M. Klusch *et al.*, The role of agents in distributed data mining: issues and benefits, *Proc. IAT03*, 2003, pp. 211–217.
16. L. Lin and L. Cao, Mining in-depth patterns in stock market, *Int. J. Intell. Syst. Technol. Appl.* 2006 (to appear).
17. B. Liu, W. Hsu, S. Chen and Y. Ma, Analyzing subjective interestingness of association rules, *IEEE Intell. Syst.* **15**(5) (2000) 47–55.
18. E. Omiecinski, Alternative interest measures for mining associations, *IEEE Trans. Knowl. Data Engin.* **15** (2003) 57–69.
19. B. Padmanabhan and A. Tuzhilin, A belief-driven method for discovering unexpected patterns, *KDD-98*, 1998, pp. 94–100.
20. A. Silberschatz and A. Tuzhilin, What makes patterns interesting in knowledge discovery systems, *IEEE Trans. Knowl. Data Engin.* **8**(6) (1996) 970–974.
21. A. Silberschatz and A. Tuzhilin, On subjective measures of interestingness in knowledge discovery, *Knowl. Discov. Data Min.* 1995, pp. 275–281.
22. P. Tan, V. Kumar and J. Srivastava, Selecting the right interestingness measure for association patterns, *SIGKDD*, 2002, pp. 32–41.
23. D. Taniar and J. W. Rahayu, Chapter 13: Parallel data mining, *Data Mining: A Heuristic Approach*, eds. H. A. Abbass, R. Sarker and C. Newton (Idea Group Publishing, 2002), pp. 261–289.
24. K. Wang, S. Zhou and J. Han, Profit mining: From patterns to actions, *EBDT*, 2002.
25. S. Yoon, L. Henschen, E. Park and S. Makki, Using domain knowledge in knowledge discovery, *Proc. Eighth Int. Conf. Information and Knowledge Management* (ACM Press, 1999).



Longbing Cao is a IEEE Senior Member, received the Ph.D. in complex systems and intelligence Sciences from Chinese Academy of Sciences, and the Ph.D. in computer science from University of Technology, Sydney. Currently, he is with Faculty of Information Technology, University of Technology, Sydney,Australia.

Chengqi Zhang is a IEEE Senior Member, received the Ph.D. in computer science from the University of Queensland, and D.Sc. degree in computer science from Deakin University. Currently, he is an research professor with Faculty of Information Technology, University of Technology, Sydney, Australia.