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 regarding a business issue. Driven by this methodology, domain intelligence is not necessary in targeting the demonstration of an algorithm. As a result, very often knowledge discovered is not generally interesting to business needs. However, real-world applications expect knowledge for taking effective actions. To this end, this paper proposes domain-driven data mining methodology, which involves domain intelligence into mining actionable knowledge in constrained environment for satisfying user needs. Key components of domain-driven data mining are constrained context, integrating domain intelligence, human-machine cooperation, in-depth mining, actionability enhancement, and iterative refinement process. We illustrate two case studies of utilizing domain-driven data mining methodology: mining impact-targeted activity patterns and identifying stock trading patterns of interest to trading. The results show that domain-driven data mining has a potential for further enhancing the actionability of mined patterns in real-world situation.

Keywords: domain-driven data mining, actionable knowledge discovery, domain intelligence

1. Introduction

In the last ten years, data mining, or KDD (knowledge discovery in database) (Han et al 2006), has been an active research and development area in existing information technology fields. In particular, data mining is gaining rapid development in comprehensive aspects such as data analyzed, knowledge discovered, techniques developed, and applications involved. The following Table 1 illustrates such key research and development progress in KDD.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Key research progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mined</td>
<td>• Relational, data warehouse, transactional, object-relational, active, spatial, time-series, heterogeneous, legacy, WWW</td>
</tr>
<tr>
<td></td>
<td>• Stream, spatiotemporal, multi-media, ontology, event, activity, links, graph, text, etc.</td>
</tr>
<tr>
<td>Knowledge discovered</td>
<td>• Characters, associations, classes, clusters, discrimination, trend, deviation, outliers, etc.</td>
</tr>
<tr>
<td></td>
<td>• Multiple and integrated functions, mining at multiple levels, exceptional, etc.</td>
</tr>
<tr>
<td>Techniques developed</td>
<td>• Database-oriented, association and frequent pattern analysis, multidimensional and OLAP analysis methods, classification, cluster analysis, outlier detection, machine learning, statistics, visualization, etc.</td>
</tr>
<tr>
<td></td>
<td>• 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.</td>
</tr>
<tr>
<td>Application involved</td>
<td>• Engineering, retail market, telecommunication, banking, fraud detection, intrusion detection, stock market, etc.;</td>
</tr>
<tr>
<td></td>
<td>• Specific task-oriented mining</td>
</tr>
<tr>
<td></td>
<td>• Biological, social network analysis, intelligence and security, etc.</td>
</tr>
<tr>
<td></td>
<td>• Enterprise data mining, cross-organization mining</td>
</tr>
</tbody>
</table>

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

Many researchers and developers have realized the limitation of extant data mining methodologies and
approaches, and the gap between business interestingness and academic attention. 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 (Ankerst 2002, Fayyad et al 2003), 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 outcomes of expectation to business. However, it is not a trivial task to deliver actionable knowledge by existing KDD approaches. This situation partly results from the scenario that extant data mining is a data-driven trial-and-error process (Ankerst 2002), 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 conditions and expectation of mined results, for instance, financial data mining and crime pattern mining is highly constraint-based (Boulicaut et al 2005, Fayyad 2003). The difference gets involved in key aspects such as problem concerned, KDD context, mined, patterns interested, processes of mining, interestingness cared, and infrastructure supporting data mining.

To handle the above difference, our experience (Cao and Dai 2003a and 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 together, a sleek data mining methodology can be developed 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 the following key components (i) problem understanding and 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 presents 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 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 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 refines the lifecycle of data mining in an iterative manner.

Further, we illustrate three case studies of domain-driven data mining in the real world. They are domain-driven stock mining, impact-targeted activity mining in security-related areas, and Web visitor classification. These instances demonstrate that domain-driven data mining can benefit actionable knowledge mining in the real world in a more effective and efficient manner than usual data-driven methodology such as CRISP-DM (CRISP).

The remainder of this paper is organized as follows. Section 2 discusses the evolution of KDD from data-driven to domain-driven. Section 3 presents major criteria for measuring the actionability of knowledge. In section 4, key components in domain-driven data mining are stated. Section 5 introduces a domain-driven data mining framework. Two case studies utilizing domain-driven data mining methodology are demonstrated in Section 6. We conclude this paper and present future work in Section 7.

2. KDD: Data Driven vs. Domain Driven

One of the fundamental objectives of KDD is to discover knowledge of main interest to real business needs and user preference. However, this forms a big challenge to extant and future data mining research and applications. To better understand this conflict, we need go back to 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 the 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 has just or mainly been 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 must 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 extend 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 bigger than a threshold presumed by data mining researchers then the algorithm is promising, (viii) among (vii) the accuracy of an algorithm is taken as a key criteria of quality judgment.

In a summary, traditional KDD is a data-driven trial-and-error process targeting automated hidden knowledge discovery (Ankerst 2002, Cao & Zhang 2006). The goal of traditional data mining is to let data to create/verify research innovation, demonstrate and push the use of novel algorithms discovering knowledge of interest to researchers.

2.2. Real world KDD expectation: 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 (Ankerst 2002, Cao et al 2006b) and retrospective literature. This situation partly results from the limitation of extant data mining methodologies, which do not take into much consideration of 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. While 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 three key points have recently been highlighted: comprehensive constraints around a problem (Boulicaut et al 2005), domain knowledge and human role (Ankerst 2002, Han 1999, Cao & Dai 2003a) 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 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 at 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 (Cao & Zhang 2006). The basic theory of domain-driven data mining is as follows. On top of the data-driven framework, it aims to developing proper methodologies and techniques for integrating domain knowledge, human role and interaction, as well as actionability measures into the KDD process, which target to discover actionable knowledge in a practical 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 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 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.

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Traditional data-driven</th>
<th>Domain-driven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object mined</td>
<td>Data tells the story</td>
<td>Data and domain (business rules, factors etc.) tell the story</td>
</tr>
<tr>
<td>Aim</td>
<td>Developing innovative approaches</td>
<td>Generating business impacts</td>
</tr>
<tr>
<td>Objective</td>
<td>Algorithms are the focus</td>
<td>Systems are the target</td>
</tr>
</tbody>
</table>

Table 2. Data-driven vs. domain-driven data mining
### Technical Objective Measures

Measures of interest specified for a data mining method.

### Definition 1.

Discovered in association rule $P$ is technically interesting if it satisfies statistical significance. It could be a set of criteria. For instance, the following logic formula indicates that an association rule $P$ is interesting if it satisfies technical objective measures:

$$
\forall x \in I, \exists P : x.\text{min}\_\text{support}(P) \land x.\text{min}\_\text{confidence}(P) \Rightarrow x.\text{tech}_\text{obj}(P)
$$

### Definition 2.

Let data to create/verify research innovation; demonstrate and push the use of novel algorithms discovering knowledge of interest to research. Let data and domain knowledge to 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 non-actionable to real needs due to interestingness gaps between academia and business (Gur et al 1997). Therefore, it is critical to get a clear answer to the problem “what makes KDD of interest to business”. 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 (Tan et al 2002). 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 on not only technical measures but also 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 interestingness measurement of data-driven vs. domain-driven data mining.

### Integrative - Actionability

<table>
<thead>
<tr>
<th>Objective</th>
<th>Technical objective $tech_{obj}$</th>
<th>Technical subjective $tech_{subj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Subjective</td>
<td>Business subjective $biz_{subj}$</td>
</tr>
<tr>
<td>Integrative</td>
<td>-</td>
<td>Actionability $act$</td>
</tr>
</tbody>
</table>

### Table 3. Interestingness measurement of data-driven vs. domain-driven data mining

In the following, we give definitions of the interestingness measurement. 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, $x$ is 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 $tech_{int}$() of a rule or a pattern is highly dependent on certain technical measures of interest specified for a data mining method. Technical interestingness is further measured in terms of technical objective measures $tech_{obj}$() and technical subjective measures $tech_{subj}$().

**Definition 2.** Technical objective interestingness $tech_{obj}$() captures 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 $\text{min}_\text{support}$ and $\text{min}_\text{confidence}$.

$$
\forall x \in I, \exists P : x.\text{min}\_\text{support}(P) \land x.\text{min}\_\text{confidence}(P) \Rightarrow x.\text{tech}_\text{obj}(P)
$$

**Definition 3.** Technical subjective interestingness $tech_{subj}$() also focuses and is based on technical means, recognize to what extent a pattern is of interest to a particular user needs. For instance, probability-based belief (Padmanabhan et al 1998) is developed for measuring the expectedness of a link pattern.

**Definition 4.** 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 is also represented by a collection of criteria from both objective $biz_{obj}$() and subjective $biz_{subj}$() perspectives.

**Definition 5.** Business objective interestingness $biz_{obj}$() measures to what extent that 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) is often 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.

\[ \forall x \in I, \exists P: x.prof(P) \land x.roi(P) \Rightarrow x.biz_obj(P) \]  

**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 \( psycho-index \) may be used to indicate that a trader thinks it as 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 \( P \), its actionable capability \( act() \), is described as to what degree that it can satisfy both the technical and the business interestingness.

\[ \forall x \in I, \exists P: act(P) = f(tech_{obj}(P) \land tech_{subj}(P) \land biz_{obj}(P) \land biz_{subj}(P)) \]  

If a pattern is automatically discovered by a data mining model while it only satisfies technical interestingness request, it is usually 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.

\[ \forall x \in I, \exists P: x.tech_int(P) \land x.biz_int(P) \Rightarrow x.act(P) \]  

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

To illustrate the above theory, we present an example of measuring the interestingness of mining activity sequences in social security transactions which can indicate the probability of associating with government customer debt. To this end, we develop a series of new technical and business interestingness metrics, and use sequential association rule mining to find activity sequences, and develop business interestingness metrics based on debt duration, debt amounts, etc. Section 6.1 presents some details.

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 data mining handling stream data, link mining studying linkage across entities. Challenges and prospects coming from the real world force us to rethink of some key points in data mining. This includes problem understanding and definition, KDD context, patterns mined, mining process, interestingness 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 vs. domain-specific**

In traditional data mining studies, researchers pay a large amount of time to construct 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 always is abstracted and pruned into a very general and brilliant research issue to fill in innovation and significance requirements of research. 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, the objectives and goals of applying KDD are basically problem-solving and 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 non-functional concerns. The analysis and modeling of these requirements request domain intelligence, namely domain background knowledge, and the 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 of 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 one of proper approaches (Cao et al 2006a) to model domain knowledge. Ontology-based specifications build a business ontological domain to represent domain knowledge in terms of ontological items and semantic relationships. We can develop ontological representations to manage the above items and relationships.

Through ontology-based representation and transformation, business terms are mapped to data mining system’s internal ontologies. So 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 will be 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 vs. constrained

Law, business rule and regulation are common forms of constraints in human society. Similarly, data mining targeting 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, interestingness constraint and deployment constraint.

The real-world business problems and requirements are often tightly embedded in domain-specific business process and business rules in charge with expertise (domain constraint). Potential matters to satisfy or react on domain constraints could consist of building domain model, domain metadata, semantics and ontologies (Cao et al 2006a), 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 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.

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). 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 are usually used by
traders to judge whether a trading rule is interesting enough or not.

Furthermore, interesting patterns often cannot be deployed to real life if they are not integrated with business rules, regulations 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 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 alarming should be minimized.

Some other types of constraints include knowledge type constraint, dimension/level constraint and rule constraint (Han 1999). Several types of constraints play significant roles in a process effectively discovering knowledge actionable to business world. 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 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 (Boulicaut et al 2005, Han 1999).

4.3. Pattern: generic vs. actionable patterns

Many mined patterns are more useful to data miners than to business persons. Generally interesting patterns are useful because they satisfy technical interestingness measurement. For instance, a large number of association rules are often found, even though most of them might not be workable in business. These rules are generic patterns or technically interesting rules.

However, they are not necessarily useful for solving business problems. To improve this situation, we advocate in-depth pattern mining which aims to developing patterns actionable in business world. It targets the discovery of actionable patterns to support smart and effective decision-making, namely a pattern must satisfy $\forall P: x.\text{tech\_int}(P) \land x.\text{biz\_int}(P) \rightarrow x.\text{act}(P)$. Therefore, in-depth patterns can be delivered through improving either technical interestingness $\text{tech\_int}(\cdot)$ or business interestingness $\text{biz\_int}(\cdot)$. As discussed in Section 3 on pattern interestingness, both technical and business interestingness measures must be satisfied from both objective and subjective perspectives.

Technically, it could be through enhancing or generating more effective interestingness measures (Omiecinski 2003), for instance, a series of research have been done on designing right interestingness measures for association rule mining (Tan et al 2002). It could 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, business interestingness needs to be checked and emphasized 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 a major role in enhancing pattern actionability. 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 mechanism through embedding domain knowledge and human involvement.

4.4. Infrastructure: automated vs. human-mining-cooperated

Traditional data mining is an automated trial and error process. Deliverables of data mining include 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 in aspects such as lacking of capability in
involving domain intelligence and adapting to dynamic situations in business world. In particular, automated data mining has big trouble in handling enterprise data mining applications.

The requirements of discovering actionable knowledge 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 likely presents as a human-machine-cooperated interactive knowledge discovery process.

The role of human 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 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 can largely improve the effectiveness and efficiency of mining actionable knowledge.

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

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 technique, multi-modal, mobile agents, 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

The existing data mining methodology, for instance 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 naive 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. 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 for dealing with multiform constraints and domain knowledge. This section outlines key ideas and relevant research issues of DDID-PD.

5.1. Process Model

The main functional components of the DDID-PD are shown in Figure 1, where we highlight those processes specific to DDID-PD in thickened boxes. The lifecycle 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 interaction with real users or domain experts.

The lifecycle 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;
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

(i) **constraint-based context**, actionable pattern discovery are based on deep understanding of the constrained environment surrounding the domain problem, data and its analysis objectives,

(ii) **integrating domain knowledge**, real-world data applications inevitably involve domain and background knowledge which is very significant for actionable knowledge discovery,

(iii) **cooperation between human and data mining system**, the integration of human role, and the interaction and cooperation between domain experts and mining system in the whole process are important for effective mining execution,

(iv) **in-depth mining**, another round of mining on the first-round results may be necessary for searching patterns really interesting to business,

(v) **enhancing knowledge actionability**, based on the knowledge actionability measures, further enhance the actionable capability of findings from modeling and evaluation perspectives,

(vi) **loop-closed iterative refinement**, patterns actionable for smart business decision-making would in most case be discovered through loop-closed iterative refinement, and

(vii) **interactive and parallel mining supports**, developing business-friendly system supports for human-mining interaction and parallel mining for complex data mining applications.

The following section outlines each of them respectively.

### 5.2. 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, integrating domain knowledge, should be paid special attention into 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, the following Figure 2 illustrates the reference model for actionability enhancement.

In the field of developing real-world data mining applications, questionnaires are very helpful for capturing
business requirements, constraints, requests from organization and management, risk and contingency plans, expected representation of the deliverables, etc. 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 into the knowledge management base for well organizing the knowledge and the process of domain-driven data mining applications.

![Diagram of the process](image)

**Fig. 2. Actionability enhancement**

### 5.3. System supports

To support domain-driven data mining, it is significant to develop interactive mining supports for human-mining interaction and evaluate the findings. On the other hand, parallel mining supports are often 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 2006a) 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 are then mapped to mining domain for mining purposes. Domain experts can help train, supervise and evaluate the outcomes.

Parallel KDD (Domingos 2003, Taniar et al 2002) supports involve parallel computing and management supports to deal with multiple sources, parallel I/O, parallel algorithms and memory storage. For instance, to tackle cross-organization transactions, we can design efficient parallel KDD computing and system supports to wrap the 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 to find good parallel computing strategies.

The facilities for interactive and parallel mining supports can largely 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, etc. They are some essential parts of next-generation KDD infrastructure.
6. Case Study

In this section, we illustrate some of our work in developing domain-driven data mining. The first example is impact-targeted activity mining in security areas. It targets mining high impact activities which likely lead to threats to national and homeland security. We demonstrate real work in social security area. The second one is to discover actionable trading strategies in generally interesting pattern set (Cao et al 2006c, Lin & Cao 2006). For space limit, we only highlight some of key components in utilizing domain-driven data mining methodology.

6.1. Impact-targeted activity mining in security areas

The domain-driven data mining theory has been used in mining impact-targeted activity patterns in social security area (Cao et al 2006d). For instance, in frequent activity sequence mining, we first 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: debt-related activity set and non-debt related activity set. To handle such unbalanced data, we develop both technical and business metrics for measuring the actionability of a pattern. For instance, the following technical metrics are defined: global support, local support, class difference rate, relative risk ratio.

**Definition 8.** The global support of a pattern \( \{P \rightarrow \$ \} \) in activity set \( A \) is defined as \( \text{Supp}_A(P, \$) = |P, A|/|A| \).

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

**Definition 9.** The local support (\( L\text{-SUPP} \)) of a rule \( \{P \rightarrow \$ \} \) in target activity set \( D \) is defined as \( \text{Supp}_D(P, \$) = |P, D|/|D| \). On the other hand, the local support of rule \( \{P \rightarrow \$ \} \) in activity set \( A-D \) (i.e., non-debt activity set) is defined as \( \text{Supp}_{A-D}(P, \$) = |P, A-D|/|A-D| \). The class difference rate \( Cdr(P, P_{D-A}) \) of \( P \) in two independent classes \( D \) and \( A-D \) is defined as

\[
Cdr(P, P_{D-A}) = \frac{\text{Supp}_D(P, \$)}{\text{Supp}_{A-D}(P, \$)}.
\]

If \( Cdr(P, P_{D-A}) \) is larger than a given threshold, then \( P \) far more frequently leads to debt than result in non-debt. 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 10.** Given local support (\( \text{SUPP} \)) \( \text{Supp}_D(P, \$) \) and \( \text{Supp}_{A-D}(P, \$) \), the relative risk ratio \( Rrr(P, P_{D-A}) \) of \( P \) leading to target activity classes \( D \) and non-target class \( A-D \) is defined as

\[
Rrr(P, P_{D-A}) = \frac{\text{Prob}(P, D)/\text{Prob}(P, \$)}{\text{Prob}(P, A-D)/\text{Prob}(P, \$)} = \frac{\text{Supp}_D(P, \$)}{\text{Supp}_{A-D}(P, \$)}.
\]

If \( Rrr(P, P_{D-A}) \) is larger than a given threshold, then \( P \) far more frequently leads to debt than results in non-debt. This measure indicates the statistical difference of a sequence \( P \) leading to debt or non-debt in a global manner. An obvious difference between them is expected for positive frequent impact-targeted activity patterns.

A number of sequential activity patterns are mined based on the above and traditional measures such as left side support (\( \text{LSUPP} \)), right hand support (\( \text{RSUPP} \)), left hand count (\( \text{LCNT} \)), right hand count (\( \text{RCNT} \)), confidence (\( \text{CONF} \)), lift (\( \text{LIFT} \)) and z score (\( \text{ZSCORE} \)). For instance, the following Table 4 illustrates one sequential activity pattern (\( \text{ADV, EVN} \rightarrow \text{DET} \)) likely associated with debt in balanced mix data.

<table>
<thead>
<tr>
<th>PATTERN</th>
<th>LSUPP</th>
<th>RSUPP</th>
<th>SUPP</th>
<th>L_SUP</th>
<th>LCNT</th>
<th>RCNT</th>
<th>CNT</th>
<th>CONF</th>
<th>LIFT</th>
<th>ZSCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV, EAN ( \rightarrow ) DET</td>
<td>0.0186</td>
<td>0.5</td>
<td>0.0157</td>
<td>0.031</td>
<td>57</td>
<td>1559</td>
<td>49</td>
<td>0.845</td>
<td>1.69</td>
<td>3.73</td>
</tr>
</tbody>
</table>
We then prune this pattern set by developing business interestingness metrics, for instance, the following ones 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 11.** The total debt amount $d_{amt}(I)$ is the sum of all individual debt amounts $d_{amt}(I) = 1, \ldots, f$ in $f$ itemsets holding the pattern ACB. Then we get pattern average debt amount $\bar{d}_{amt}$ for the pattern ACB as:

$$\bar{d}_{amt} = \frac{1}{f} \sum_{i=1}^{f} d_{amt}(I)$$

**Definition 12.** Debt duration $d_{dur}(I)$ for the pattern ACB is the average duration of all individual debt durations in $f$ itemsets holding the pattern ACB. Debt duration $d_{dur}(I)$ of an activity is the number of days a debt keeps valid, $d_{dur}(I) = d.end_date - d.start_date + 1$, where $d.end_date$ is the day a debt is completed, while $d.start_date$ is the day a debt is activated. Pattern average debt duration $\bar{d}_{dur}$ is defined as:

$$\bar{d}_{dur} = \frac{1}{f} \sum_{i=1}^{f} d_{dur}(I)$$

For instance, the following Table 5 lists technical and business interestingness measures of the activity sequence rule “LET, ANO --> DET”: for Australian Centrelink NewStart benefit recipients. If the activity “Annotation” follows “NSS Letter” in customer contacts, then this customer likely leads to government customer debt. The technical interestingness tells users the statistical significance of this rule, while business interestingness shows Centrelink officers how important this rule leads to debt cost to Centrelink.

- **Technical interestingness:**
  - support = 0.01251
  - count = 39, the number of rules triggered in the test set
  - confidence = 0.60935
  - lift = 1.2187

- **Business interestingness:**
  - debt_amt_sum = 1,151,551, the sum of debt amount in cents of those debt-related activity sequences supporting the rule in three month
  - debt_dur_sum = 605, the sum of debt duration in days of those debt-related activity sequences supporting the rule
  - debt_amt_avg = 29,526, the averaged debt amount in cents of those debt-related activity sequences supporting the rule
  - debt_dur_avg = 15.5, the averaged debt duration in days of those debt-related activity sequences supporting the rule

<table>
<thead>
<tr>
<th>Item1</th>
<th>Supp</th>
<th>Cnt</th>
<th>Conf</th>
<th>Lift</th>
<th>debt_amt_sum</th>
<th>debt_dur_sum</th>
<th>debt_amt_avg</th>
<th>debt_dur_avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>LET, ANO --&gt; DET</td>
<td>0.01251</td>
<td>39</td>
<td>0.60935</td>
<td>1.2187</td>
<td>1,151,551</td>
<td>605</td>
<td>29,526</td>
<td>15.5</td>
</tr>
</tbody>
</table>

6.2. Domain-driven stock data mining

Financial data mining (Kovalerchuk et al 2000) is of high interest since it may benefit trading decision and market surveillance, but also challenging because financial markets are greatly complex. Taking ASX as an instance, there are more than 1000 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 illustrates some results of the above work in ASX data.

In order to support actionable trading pattern mining, we define the following metrics: Support, Confidence, All_Confidence, Cosine and Coherence are defined for measuring the actionability of trained trading evidences in
in-sample set when they are tested in out-of-sample data. In our situation, support equals to coherence.

**Definition** 13. Let $D$ be the number of total trading evidences (e.g., rule-stock pairs) found in in-sample and out-of-sample sets satisfying certain business interestingness (say return larger than a threshold $R_0$). $A$ ($B$) be the number of total evidences in in-sample (out-of-sample) data set. $AB$ be the number of pairs existing in both in-sample and out-of-sample sets, which satisfying business interestingness request. The following probability functions are defined for trading evidences in or across in-sample and out-of-sample data sets:

$$P(A) = \frac{A}{D}, \quad P(B) = \frac{B}{D}, \quad P(AB) = \frac{AB}{D}, \quad P(A|B) = \frac{P(AB)}{P(B)}, \quad P(B|A) = \frac{P(AB)}{P(A)}.$$ Further, we define the following metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support $P(AB)$</td>
<td>$\max(\frac{A}{D}, \frac{B}{D})$</td>
</tr>
<tr>
<td>Confidence</td>
<td>$\frac{AB}{D}$</td>
</tr>
<tr>
<td>All Confidence</td>
<td>$\frac{P(AB)}{\max(P(A), P(B))}$</td>
</tr>
<tr>
<td>Cosine</td>
<td>$\frac{P(AB)}{\sqrt{P(A)P(B)}}$</td>
</tr>
<tr>
<td>Coherence</td>
<td>$\frac{P(AB) - P(A) - P(B)}{P(A) + P(B) - P(AB)}$</td>
</tr>
</tbody>
</table>

In general, the value range of the above metrics is in the interval of [0, 1]. Larger value donates bigger actionability of trading evidences when they are deployed into the real trading. Taking the trading rule-stock pair mining as an instance, the following Table 6 provides the summary statistics of actionability for the discovered rule-stock pairs from April to October 2001 using ASX intraday orderbook data. We take sliding window strategy for training and testing, i.e., use month $a$ data as in-sample, the data of the next month $a+1$ as out-of-sample. After this round of training-testing, month $a+1$ will be the in-sample, and the out-of-sample is month $a+2$.

When we specify the top $x\%$ as benchmark for selecting pairs from in-sample and out-of-sample, the statistics of both in-sample and out-of-sample are same. In this case, there are the following relations existing among the above five metrics.

$$Confidence = All\ _Confidence = Consine$$

Further, varied financial functions are specified for measuring and pruning the interestingness of identified general trading patterns, for instance, *profit*, *return* and *sharpe ratio*. The following defines *sharpe ratio* $SR$ used in stock market.

**Definition** 14. The *sharpe ratio* of a trading pattern is:

$$SR = \frac{(R_p - R_f)}{\sigma_p}$$

Where $R_p$ is expected portfolio return, $R_f$ is risk free rate, $\sigma_p$ is portfolio standard deviation.

*Sharpe ratio* measures the performance of a trading rule in terms of both return and risk. If *sharpe ratio* is high, the rule likely leads to high return with low risk.

Based on the above strategies, we identify highly correlated association between trading rules and stocks in Australian Stock Exchange orderbook data in 2001. Table 6 illustrates the summary statistics for $SR$ larger than or equal to 0.4. In this case, we find the following frequent rule-stock pairs: Rule 1-Stock 14, Rule 1-Stock 24, Rule 2-Stock 14, Rule 2-Stock 18, Rule 2-Stock 24 (association support $\geq 20\%$).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sept</th>
<th>Oct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>14.8%</td>
<td>21.1%</td>
<td>16.7%</td>
<td>16.7%</td>
<td>20%</td>
<td>16.7%</td>
<td>16.7%</td>
</tr>
<tr>
<td>Confidence</td>
<td>40%</td>
<td>40%</td>
<td>50%</td>
<td>50%</td>
<td>41.7%</td>
<td>28.6%</td>
<td>33.3%</td>
</tr>
<tr>
<td>All Confidence</td>
<td>21.1%</td>
<td>26.7%</td>
<td>20%</td>
<td>20%</td>
<td>27.8%</td>
<td>28.6%</td>
<td>25%</td>
</tr>
<tr>
<td>Consine</td>
<td>27.6%</td>
<td>32.7%</td>
<td>31.6%</td>
<td>31.6%</td>
<td>34%</td>
<td>28.6%</td>
<td>28.9%</td>
</tr>
</tbody>
</table>

Figure 3 further compares the market index return with our monthly return of top 5% pairs (using sliding window strategy) in ASX 2001 orderbook data after deducting transaction cost. This result shows that in ASX bear
market situation of 2001, our strategy can still beat the market index return, which is very undulate. This almost positive result demonstrates that trading on these pairs can beat the market. It demonstrates that the discovered rule-stock pairs are promising for the real trading.

7. 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 a potential to change the existing situation 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.

This paper has proposed a new data mining methodology, referred to as domain-driven data mining, based on the retrospective review of traditional data mining. It provides a systematic overview of the issues in discovering actionable knowledge, and advocates the methodology of mining actionable knowledge in constrained 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 include almost all phases of the well-known industrial data mining methodology CRISP-DM. However, it also highlights some missing key points in traditional data-driven methodologies, for instance

- mining in-depth patterns,
- involving domain intelligence, and
- developing both subjective and objective measures from technical and business perspectives.

These components play key roles in improving the existing knowledge discovery towards a more effective and practical manner. It has a potential of delivering a packed solution satisfying real user needs rather than demonstrating specific algorithms.

The domain-driven data mining methodology is used in telecom data mining, financial data mining, and government service mining. We have demonstrated some of our research results in mining impact-targeted activity sequences in security areas, and actionable trading patterns in stock markets. The experiments show that domain-driven data mining has potential for improving the actionable knowledge mining. Our current work is on developing detailed domain-driven data mining specifications, interfaces and infrastructure supports for easily deploying domain-driven data mining methodology into real-world mining.

Acknowledgments
Data used in this paper is from Australian Centrelink and Capital Market CRC, AC3 and SIRCA. We are grateful to the support and business knowledge from related organizations, for instance, Mr Rick Schurmann from Centrelink, and Prof. Mike Aikten from CMRC.

References

19. J. Han, “Towards Human-Centered, Constraint-Based, Multi-Dimensional Data Mining,” An invited talk at Univ. Minnesota, Minneapolis, Minnesota, 1999.