Domain-Driven Actionable Knowledge Discovery in the Real World[*]

Longbing Cao and Chengqi Zhang

Faculty of Information Technology, University of Technology Sydney, Australia

Abstract. Actionable knowledge discovery is one of Grand Challenges in KDD. To this end, many methodologies have been developed. However, they either view data mining as an autonomous data-driven trial-and-error process, or only analyze the issues in an isolated and case-by-case manner. As a result, the knowledge discovered is often not actionable to constrained business. This paper proposes a practical perspective, referred to as *domain-driven in-depth pattern discovery* (DDID-PD). It presents a domain-driven view of discovering knowledge satisfying real business needs. Its main ideas include constraint mining, in-depth mining, human-cooperated mining, and loop-closed mining. We demonstrate its deployment in mining actionable trading strategies in Australian Stock Exchange data.

1 Introduction

Actionable knowledge discovery can afford important grounds to business decision makers. In the panel discussions of SIGKDD 2002 and 2003 [2, 7], it was highlighted by panelists as one of the Grand Challenges for extant and future data mining. 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 via some predefined models based on experts' hypothesis. Data mining is presumed as an automated process producing automatic algorithms and tools without human involvement and the capability to adapt to external environment constraints.

However, data mining in the real world, for instance financial data mining, is highly constraint-based [8, 11]. Constraints involve technical, economic and social aspects. The real-world business problems and requirements are often tightly embedded in domain-specific business rules and process with expertise (*domain constraint*). Patterns actionable to business are often hidden in large quantities of data with complex structures, dynamics and source distribution (*data constraint*). Often mined patterns are not actionable to business even though they are interesting to research. There exist big interestingness gaps between academia and business (*interestingness constraint*). Furthermore, interesting patterns often cannot be deployed to real life if they are not integrated with business rules, regulations and processes (*deployment constraint*). There could be other types of constraints such as knowledge constraint, dimension/level constraint and rule constraint [8].

[*] This work is sponsored by UTS Chancellor research fund and ECRG Fund.

W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 821 – 830, 2006.

To discover actionable knowledge from data embedded in the above constraints, it is essential to slough off the superficial and captures the essential information from the data mining. However, this is a non-trivial task. Tricks may not only include how to find a right pattern with a right algorithm in a right manner, they also involve a suitable process-centric support with a suitable deliverable to business. Even many methodologies are studied, they either view data mining as an automated process, or deal with the constraints in a case-by-case manner. Our experience [3] and lessons learned in data mining in capital markets [6] 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, it can advise the process of real-world data mining in a manner more actionable and reliable to business. These are our motivation to develop a practical framework, called *domain-driven in-depth pattern discovery* (DDID-PD), for the discovery of actionable knowledge from the real world.

DDID-PD views actionable knowledge discovery as an iteratively interactive in-depth pattern mining process in domain-specific context. It exploits key components including (i) constraint mining, (ii) incorporating domain knowledge through human-mining-cooperation, (iii) in-depth mining, and (iv) loop-closed mining. Mining constraint-based context requests to develop workable mechanisms to deal with comprehensive constraints. The involvement of domain experts and their knowledge can reduce the complexity of the knowledge discovery process in the constrained world. In-depth pattern mining discovers actionable patterns. A system following the DDID-PD framework can embed effective supports for domain knowledge and experts' feedback, and refines the lifecycle of data mining in an iterative manner.

Taking financial data mining as an example, this paper introduces some case studies deploying the DDID-PD framework to mine actionable trading strategies for improving trading performance and costs. It shows that the DDID-PD can benefit the actionable knowledge mining in a more realistic and reliable manner than data-driven methodology such as CRISP-DM [13].

2 Domain-Driven In-Depth Pattern Discovery

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. This section outlines key ideas and relevant research issues of the DDID-PD.

2.1 Pattern Actionability

Let $I = \{i_1, i_2, . . . , i_m\}$ be a set of items, *DB* be a database consisting of a set of transactions, x is an itemset in *DB*. Let P be a pattern discovered in *DB* through a model M. In DDID-PD [4], the following concepts measure pattern actionability, i.e., whether or not, or to what extent, P can be used to answer real business needs.

DEFINITION 1. Technical Interestingness – The technical interestingness *tech_int*() measures how interesting the pattern is from technical perspective. It is measured through certain technical metrics specified for a data mining method. For instance, the following logic formula indicates that an association rule *P* is technically interesting if it satisfies a user-defined *min_support* and *min_confidence*.

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

DEFINITION 2. Business Interestingness – The business interestingness *biz_int*() of a pattern is determined by some domain-oriented social and/or economic criteria accepted by real users. In stock data mining, a stock price predictor *P* is interesting to trading if it satisfies the *profit* and *roi* (*return on investment*) requests.

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

DEFINITION 3. Actionability of a pattern – The actionability of a pattern *act*() indicates to what degree it satisfies both technical and business interestingness. If both technical and business interestingness or a hybrid interestingness measure integrating both aspects are satisfied, it is called an *actionable* pattern. Such kind of patterns are not only interesting to data miners, but generally interesting to decision-makers.

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

2.2 Actionable Knowledge Discovery Process

The components of the DDID-PD are shown in Figure 1, where we highlight those processes specific to DDID-PD in thicken 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.

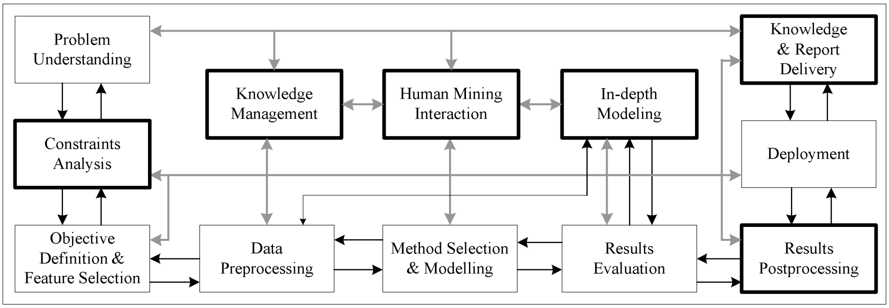


Fig. 1. DDID-PD process model

P1. Problem understanding;
P2. Constraints analysis;
P3. Analytical objective definition, feature construction;
P4. Data preprocessing;
P5. Method selection and modeling; or
P5'. In-depth modeling;
P6. Initial generic results analysis and evaluation;

P7. *It is quite possible that each phase from P1 may be iteratively reviewed through analyzing constraints and interaction with domain experts in a back-and-forth manner; or*

P7': In-depth mining on the initial generic results where applicable;

P8. Results post-processing;

P9. Reviewing phases from P1 may be required;

P10. Deployment;

P11. Knowledge delivery and report synthesis for smart decision making.

The DDID-PD process highlights four highly correlated procedures that are critical for the success of data mining in the real world. They are (i) *constraint mining*, (ii) *in-depth mining*, (iii) *human-cooperated mining*, and (iv) *loop-closed mining*. The following sections discuss them respectively.

2.3 Constraint Mining

Specifically, in Section 1, we list several types of constraints, which play significant roles in a process effectively discovering knowledge actionable to business. In practice, many other aspects such as data stream and the scalability and efficiency of algorithms may be enumerated. In DDID-PD, constraints are domain-specific, functional, nonfunctional and environmental. These ubiquitous constraints form a *constrained context* for actionable knowledge discovery. All the above constraints must, to varying degrees, be considered in relevant phases of DDID-PD. In this case, the analysis is called *constraint mining* [8, 11].

Some major aspects of domain constraints include the domain and characteristics of a problem, domain terminology, specific business process, policies and regulations, user profiling and favorite deliverables. Potential matters to satisfy or react on domain constraints could consist of building domain model, domain metadata, semantics and ontologies [5], 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.

Constraints on particular domain 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.

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. In DDID-PD, social and economic interestingness such as user preferences and domain knowledge are also considered in assessing whether a pattern is actionable or not.

Furthermore, DDID-PD advocates the delivery of an interesting pattern integrated with the domain environment such as business rules, process, information flow, presentation, etc. In addition, many other realistic issues are considered. For instance, a software infrastructure may be established to support the full lifecycle of data mining; the infrastructure needs to integrate with existing enterprise information systems and workflow; parallel KDD [10] with parallel supports are implemented on multiple sources, parallel I/O, parallel algorithms and memory storage; visualization, privacy and security should receive much-deserved attention.

2.4 In-Depth Mining

In general, data mining publications tend to push the use of specific algorithms rather than answer real business needs. As a result, patterns interesting to data miners often can not achieve business benefits when deployed. We call them *generic* patterns. Such situations have hindered the deployment and adoption of data mining in real applications. Therefore it is essential to evaluate the actionability of a pattern and focus on discovering actionable patterns satisfying both *tech_int*(*P*) and *biz_int*(*P*) to support realistic and reliable decision-making. This is *in-depth pattern mining*. Its objective is not to push the use of a specific algorithm, rather try to answer real business needs in a workable manner.

In-depth patterns mining targets to improve both technical (*tech_int()*) and business (*biz_int()*) interestingness in the above constraint-based context. Technically, it could be through enhancing or generating more effective interestingness measures [12]. It could also be through developing alternative models for discovering deeper patterns. Some other options include rule reduction, model refinement or parameter tuning by optimizing *generic* pattern set. Additionally, techniques can be developed to deeply understand, select and refine the target data set.

In in-depth mining, more attention should be paid to business requirements, objectives, domain knowledge and qualitative intelligence of domain experts for their impact on mining deep patterns. This could be through selecting and adding business features, considering domain and background knowledge in modeling, supporting interaction with domain experts, fine 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, etc.

2.5 Human Cooperated Mining

The real requirements for discovering actionable knowledge in constraint-based context determine that real data mining is more likely to be human involved rather than automated. Human involvement is embodied through cooperation between humans (including users and business analysts, mainly domain experts) and data mining system. This is achieved through the compensation 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.

In DDID-PD, the role of human (mainly domain users and experts) could 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 evaluate mining results. This assistance may largely improve the effectiveness and efficiency of mining actionable knowledge.

In general, human often serve on the feature selection and result evaluation. DDID-PD views that human could 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, human mining cooperation could be, to varying degrees, human-centred or guided mining [2, 8], or human-supported or assisted mining, etc.

To support human involvement, human mining interaction, or in a sense presented as interactive mining [1, 2], is absolutely necessary. Interaction often takes explicit form, 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 quality relies on performance such as user-friendliness, flexibility, run-time capability and understandability.

2.6 Loop-Closed Mining

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

The above key points of the DDID-PD indicate that real-world data mining cannot be dealt just with an algorithm, rather it is really necessary to build a proper data mining infrastructure to discover actionable knowledge from constraint-based scenarios in a loop-closed iterative manner. To this end, agent-based data mining infrastructure [14] presents good facilities since it provides good supports for both autonomous problem-solving and user modeling and user agent interaction.

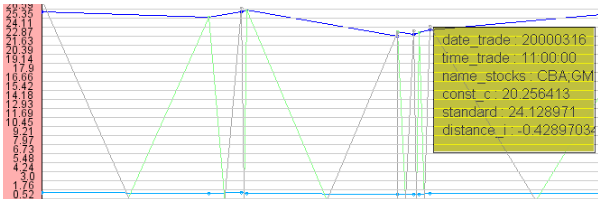
3 Case Study: Developing Actionable Trading Strategies

In stock data mining [6], we deploy the DDID-PD to mine actionable trading patterns. Our objective is to develop actionable trading strategies which can not only trigger profitable trading signals, but also result in the proper measurement and support of market dynamics. For space limit, we only illustrate two case studies here. There is other work under development in terms of domain-driven perspective such as trading strategy-stock correlation analysis, broker-based association analysis, and so forth.

3.1 Designing Actionable Trading Strategy

A quality trading strategy can be designed from scratch via analyzing market dynamics and microstructure. For instance, in real trading, traders often trade multiple stocks to manage risk. DDID-PD based technology extracts evidences about what stocks are correlated with others, and discover trading patterns effective on multiple instruments. A typical example is the pairs trading strategy, which is based on the correlation analysis between stocks. We find that an effective pairs trading strategy is not only dependent on correlations but also considering constraints and domain knowledge such as relevant market factors.

The design of an actionable strategy is a human-mining interaction process supporting iterative development, back-testing, refinement and optimization of trading strategies. To this end, we built a financial trading rule automated development and evaluation system called F-Trade[1]. Figure 2 shows signals for a stock pair in ASX market. Figure 3 further shows the impact of business factors – distance and weight on return and the number of triggered signals.



date_trade : 20000316
time_trade : 11:00:00
name_stocks : CBA;GM
const_c : 20.256413
standard : 24.128971
distance_i : -0.4289703

Fig. 2. Simulated trading via pairs trading strategy in F-Trade (ASX intraday data from 1 Jan 2000 to 20 Jun 2000)

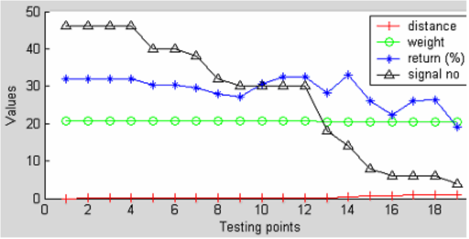


Fig. 3. Relation between d_0, *weight*, *return* and *signal number*

The exercise in testing ASX Top 32 stocks from January 1997 to June 2002 shows that DDID-PD based strategy has led to some interesting emergence beyond the normal mining algorithm design and domain expectation.

- Pair relationship between stocks and the combination of the above four factors interesting to trading cannot just be determined by technical measures such as coefficient ρ. They are also highly affected by stock movement such as volatility and liquidity. High volatility improves return while high liquidity balances the market impact on return.
- All 13 correlated stocks mined in Top 32 ASX come from different sectors. This finding means that pairs are not necessary from the same sector as presumed by financial researchers.

[1] F-Trade: accessible from http://www.f-trade.info with authorization.

3.2 Mining Actionable Trading Strategy

There exist many generic trading strategies in the literature and trading houses. Let's take the very common moving average strategy MA(*sr*, *lr*) as an instance. It actually indicates a generic correlated trading pattern between indicators *short-run moving average* (*sr*) and *long-run moving average* (*lr*).

```
ALGORITHM 1: A generic strategy MA(sr, lr)
IF $sr > lr$ THEN Buy
IF $sr < lr$ THEN Sell
```

Generally speaking, this pattern doesn't work in market trading. The actionability of a MA instance is determined in terms of the performance in real data, traders' interestingness and market dynamics such as transaction cost. To this end, the involvement of domain knowledge is quite significant for finding actionable rules. Using the DDID-PD ideas, we improve the generic MA and design an in-depth rule MA(*t*, *sr*, *lr*, δ_x, δ_y, *h*, *d*) as follows.

```
ALGORITHM 2: A revised MA($t, sr, lr, \delta_x, \delta_y, h, d$)
IF $sr*(1-\delta_x) >= lr$; triggering 'buy' signal
    $t = t+h$; holding 'h' transactions or days
    IF $sr*(1-\delta_x) >= lr$ THEN
        Buy; 'buy' signal is steady
        $t = t+d$; delaying 'd' transactions or days
IF $sr*(1+\delta_x) <= lr$; triggering 'sell' signal
    $t = t+h$; holding 'h' transactions or days
    IF $sr*(1+\delta_x) <= lr$ THEN
        Sell; 'sell' signal is steady
        $t = t+d$; delaying 'd' transactions or days
```

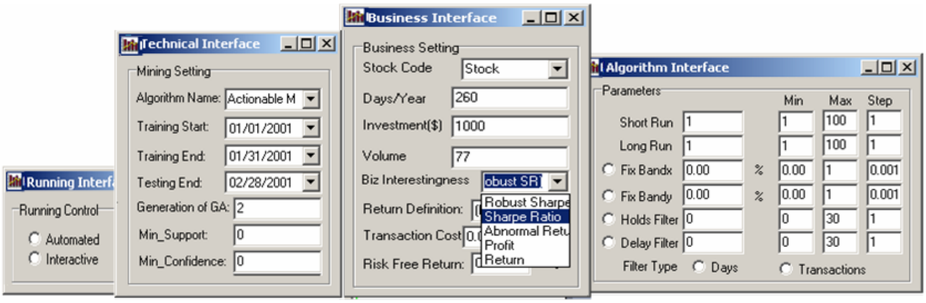
This in-depth rule considers the following constraints and background knowledge, which make it more adaptable to market dynamics compared with MA(*sr*, *lr*).

- More filters are imposed on the generic MA to assist in filtering out false trading signals which would result in losses, for instance, fixed percentage band filter δ, time delay filter *d*, and time hold filter *h*;
- The fixed band filter δ_x (or δ_y) requires the buy or sell signal to exceed *sr* or *lr* by a fixed multiplicative band δ_x (or δ_y);
- The time delay filter *d* requires the buy or sell signal to remain valid for a prespecified number of transactions or days *d* before action is taken;
- The time hold filter *h* requires the buy or sell signal to hold the long or short position for a prespecified number of transactions or days *h* to effectively ignore all other signals generated during that time;
- In practice, note that only one filter is imposed at a given time.

Furthermore, we built interaction interfaces to support the definition and refinement of both technical and business parameters. Figure 4 illustrates some of such interfaces for the revised MA(*t*, *sr*, *lr*, δ_x, δ_y, *h*, *d*). Through the interfaces, users can trigger the process in terms of either Automated execution or Interactive mode with the involvement of users. In Interactive mode, technical analysts can advise the above process as well as refining technical factors for setting data mining process and tuning algorithm parameters. Business analysts can supervise the construction of features, fine tune the parameters, and set evaluation criteria for the business concerns. For instance, the measure *sharpe_ratio* is used for evaluating the business actionability of an identified rule.

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

where r_P is expected portfolio return, r_R is risk free rate, and σ_P is portfolio standard deviation. Higher *sharpe_ratio* means more return with lower risk. Additionally, the system supports ad-hoc execution. Users can tune the parameters and interestingness measure at run time to evaluate the strategy.



Running Control
○ Automated
○ Interactive

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

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

Parameters

			Min	Max	Step
Short Run	1		1	100	1
Long Run	1		1	100	1
○ Fix Bandx	0.00	%	0.00	1	0.001
○ Fix Bandy	0.00	%	0.00	1	0.001
○ Holds Filter	0		0	30	1
○ Delay Filter	0		0	30	1

Filter Type ○ Days ○ Transactions

Fig. 4. Interfaces supporting human-mining system interaction

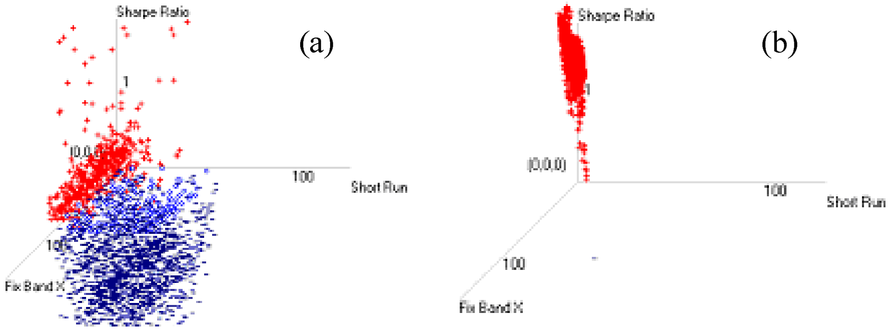


Fig. 5. Improved business interestingness by actionable rules

DDID-PD assists us in finding a collection of actionable rules. For instance, in ASX interday data, MA(4, 19, 0.033) could be an actionable rule using training data from 1 January 2000 to 31 December 2000 and testing set between 1 January 2001 and 31 December 2001. The number of trading signals generated by this rule is much bigger with better *sharpe_ratio* than other possible rules. Figure 5 (b) shows that its *sharpe_ratio* has a greatly improved positive scope compared with (a) the results of a generic MA rule. This demonstrates that DDID-PD driven strategy mining can improve strategy actionability.

4 Conclusions and Future Work

Actionable knowledge discovery is significant and also very challenging. It is nominated as one of Grand Challenges of KDD in the next 10 years. The research on this issue may change the existing situation where a great number of rules are mined while few of them are interesting to business, and promote the widely deployment of data mining into business. This paper has developed a new data mining framework, referred to as Domain-Driven In-Depth Pattern Discovery (DDID-PD). It provides a systematic overview of the issues in discovering actionable knowledge, and advocates the methodology of mining actionable knowledge in constraint-based context through human-mining system cooperation in a loop-closed iterative refinement manner.

The main phases and components of the DDID-PD include almost all phases of the CRISP-DM. It has enclosed some big differences from the CRISP-DM. For instance, (i) some new essential components, such as constraint mining, in-depth mining, the involvement of domain experts and knowledge, are taken into the lifecycle of KDD for consideration, (ii) in the DDID-PD, the normal steps of CRISP-DM are enhanced by dynamic cooperation with domain experts and the consideration of constraints and domain knowledge. These differences actually play key roles in improving the existing knowledge discovery in a more realistic and reliable way.

References

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

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

[3] Cao, L., Dai., R., Human-Computer Cooperated Intelligent Information System Based on Multi-Agents, *ACTA AUTOMATICA SINICA*, 29(1):86-94, 2003.

[4] Cao, L., et al., Domain-driven in-depth pattern discovery: a practical perspective. Proceeding of AusDM, 101-114, 2005.

[5] Cao, L., et al., Ontology-Based Integration of Business Intelligence. *Int. J. on Web Intelligence and Agent Systems*, Vol.4 No 4, 2006.

[6] Financial data mining program: http://datamining.it.uts.edu.au/.

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

[8] Han, J., Towards Human-Centered, Constraint-Based, Multi-Dimensional Data Mining. *An invited talk* at Univ. Minnesota, Minneapolis, Minnesota, Nov. 1999.

[9] Tan, P., Kumar, V., Srivastava, J., Selecting the Right Interestingness Measure for Association Patterns, SIGKDD’02, pp32-41.

[10] Manlatty,M., etc. Systems support for scalable data mining, *SIGKDD Explorations*, 2(2):56-65, 2000.

[11] J-F. Boulicaut, B. Jeudy. Constraint-based data mining. The Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach (Eds.), Springer, pp. 399-416, 2005.

[12] Omiecinski, E., Alternative Interest Measures for Mining Associations. *IEEE Transactions on Knowledge and Data Engineering*, 15:57-69, 2003.

[13] http://www.crisp-dm.org.

[14] Zhang, C., Zhang, Z., Cao, L., Agents and Data Mining: Mutual Enhancement by Integration, *LNCS 3505*, 50-61, 2005.