ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Innovative Business Using Applicable Data Mining

Appalabattula Naga Venkata Veera Bhadram[1], M. Srinivasa Chakravarthy[2] , Dr. Yalla Venkateswarlu[3]
P.G. Student, Department of Computer Science and Engineering, GIET Engineering College, Rajahmundry, Andhra Pradesh, India[1]
Assistant Professor, Department of Computer Science and Engineering, GIET Engineering College, Rajahmundry, Andhra Pradesh, India[2]
Professor and HOD, Department of Computer Science and Engineering, GIET Engineering College, Rajahmundry, Andhra Pradesh, India[3]

Abstract: To convert the problem definition, model design to applicable pattern discovery, a problem solving process is designed, which is known as Applicable Data Mining. This is designed to deliver accurate business rules which can have close integration with technical and business processes. The auxiliary model "Multisource Combined Mining based ADM" is illustrated. This generic ADM model is designed to support ADM process. This paper demonstrates a view of ADM from the stand point of technical and decision making. A case study of MSCM-ADM is demonstrated and further experiments show that this model design is flexible, practical and general to handle complex business issues and projects by deriving applicable deliverables to propose accurate decision making methods.

Keywords: Applicable Data Mining, Domain Driven Data Mining (D^3M), Multi Source Combined Mining , Technical Interestingness, Business Interestingness

I. INTRODUCTION

Information Technology hosts many domains and Database field is being the biggest of all, it has many categories in it, which are big domains in their nature. One such emerging field is Data Mining, a rapid development area provides lot of scope of innovation and research. The algorithms and patterns developed by traditional data mining processes do not possess the adoption of constraints of external environment, because of this the algorithms are getting published but few of them are transferred into real time business patterns.

In a particular database, let a customer schema defines a set of attributes for each customer where each customer is represented by one record. So, the attribute can possibly a value in a record for a given customer and the combination of such attribute and value pairs will make a pattern for a customer. By combining these pairs, simple but effective decision patterns can be achieved. We can generalize the definition of association rules to include the non-transactional data by replacing the sets of items in the traditional definition of association rules with conjunctions of (attribute=value) equalities.

For example, if a customer is having his age as 35 or above and his salary is more than $40,000 per annum, then he can have a own house and the loan can be sanctioned for him. This can be represented with (attribute=value) patterns and their combinations as follows :

$$(age \geq 35) \wedge (salary > \$40{,}000) \Rightarrow (ownhome = Yes)$$

The identified patterns from data mining algorithms and tools, handed over to business people for future decision making. Management board cannot interpret the patterns for business use (Direct mapping of patterns to business use). This fact is revealed in surveys made in the Data Mining field for business management and their applications, following the above paradigm in selected domains [1].

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Besides the dynamic environment enclosing constrains, this may result from several aspects of challenges as follows [22].

1) Many patterns mined but they are not informative to business people, often they do not know which are truly interesting and operable for businesses needs

2) Most of the identified patterns are of commonsense based or no particular interest to business needs by which management feel confused how they should consider these patterns.

3) Management does not know how to interpret the patterns and what actions can be taken on them to support business decision-making due to lack of information about the patterns generated.

A large gap [9,11,13,15] between academic deliverables and business expectations and between data miners and business analysts are informed due to above issues. So, it is critical to develop effective techniques to bridge the gap. There is a need to develop an effective, practical and generic methodologies for Applicable Data Mining (ADM). Therefore, we need to discover applicable knowledge that is much more than simply satisfying predefined technical interestingness thresholds. Such applicable knowledge is expected to be delivered in operable forms for transparent business interpretation and action taking. One of the essential ways is to develop effective approaches for discovering patterns that not only are of technical significance [7], but also satisfy business expectations [1], and further indicate the possible applications that can be explicitly taken by business people [10,2].

The traditional data mining is suffering from significant problems in satisfying business needs. For instance, the profit mining [3] to extract more applicable interestingness [10,1] and subjective interestingness patterns and enhance the interpretation of them through explanation [4] , faces many issues to transform them to best use of real business needs. The over simplification of complicated domain with business issues, the whole focus on algorithm generation and its improvement and a small amount of attention is taken care of enhancing the KDD system to handle complexities in real work applications, is the critical issue. The critical elements in real world applications such as environment, expert knowledge and operability should be catered by the fundamental work on ADM. ADM must cater for domain knowledge [12] and balance technical significance , expectations from both objective and subjective perspectives [1], and support automatically converting patterns into deliverables in business-friendly and operable forms such as actions or rules.

For business people to interpret, validate, and action, the ADM deliverables should be business friendly and they can be seamlessly embedded into business processes and systems. Data mining has good potential to lead to gain higher productivity, decision making and smarter operation in business intelligence. Such efforts actually aim at the paradigm shift from traditionally technical interestingness oriented and data centred hidden pattern mining towards business use-oriented and domain-driven applicable knowledge discovery [9]. This is nothing but the Data Mining paradigm shift. The preliminary work on AKD mainly addresses specific algorithms and tools for the filtration, summarization and post processing of learned rules. A general AKD framework that can cater for critical elements those can also be instantiated into different approaches for many domain problems. There is very limited research work has been reported in this regard.

This paper identifies the definition and development of several AKD frameworks from the system point of view which follow the methodology of Domain-Driven Data Mining (DDDM, aka D^3M in short). Much focus is given on introducing principles, processes and concepts that are new, flexible and practical effective to AKD. Such frameworks are necessary and useful for implementing real-world data mining processes and systems.

The contributions are:

1) AKD problem from system and micro economy perspectives is redefined as to define fundamental concepts of action ability and applicable patterns

2) Defining knowledge applicable by highlighting both technical significance and business expectations that need to be considered, aggregated and balanced in AKD

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

3) A four general frameworks model to facilitate AKD is proposed

4) Demonstrating the effectiveness and flexibility of the proposed frameworks in handling real-life AKD.

The main ideas of D^3M-based AKD and the four frameworks are as follows: Table 1 lists key concepts and their abbreviations used in this project.

Table 1

Notations	Explanations
ADM	Applicable Data Mining Rule
MSCM-ADM	Multi-source + Combined mining based ADM
P	$P=\{p_1, ...p_u)$ is a pattern set
$\tilde{P}$	$\tilde{P}=\{\tilde{p}_1, ...\tilde{p}_u)$ is an applicable pattern set
$\tilde{R}$	$\tilde{R}=\{\tilde{r}_1,\tilde{r}_2,.....\}$ is a business rule set
Int (p)	Pattern p's interestingness
act(p)	Pattern p's applicability

MSCM-ADM: This takes care of ADM in either multiple data sources or large quantities of data. One data set will be considered for mining initial patterns and some learned patterns are then selected to guide further construction and pattern mining on the next data set(s). When all datasets are mined, iterative mining stops and corresponding patterns are merged / summarized into applicable deliverables.

II. RELATED WORK

For smart business functions and decision making practises, the Applicable Knowledge Discovery (AKD) is critical in promoting and releasing the productivity of data mining and knowledge discovery. The panellists of SIGKDD and ICDM earmarked the challenges in developing next generation KDD methodologies [10,14]. The significance of the term "Applicability" counts the ability of a pattern to suggest a user to take some definite actions to his benefit in the real world. This counts on the ability to give some direction to business decision making actions. The existing effective interestingness metrics are, based on developing and refining objective technical interestingness metrics Int (to( )) [17,15,6,7,8], where the complexities of patter structure and statistical significance are aimed in this methodology. Other work is subjective technical interestingness measures Int (t_s()) [17, 16, 18], which also recognise to what extent a pattern is of interest to participate user preferences. A probability-based belief would have been used to describe user confidence of unexpected rules [17], to say as an example.

Very limited research on developing business-oriented interestingness exists at the moment. One example among few is profit mining [3]. Limitations for the existing work on interestingness development lie in a number of aspects. Much work is on developing alternative interest measures focusing on technical interestingness [19]. The emerging research on general business-oriented interestingness is isolated from technical significance. Knowledge applicability needs to pay equal attention to both technical and business-oriented interestingness from both objective and subjective perspectives [1], would be the best solution to possible questions like what makes interesting patterns applicable in the real world. With ADM approach, the existing work focuses on developing post analysis techniques to filter or prune rules [4], reduce redundancy [20], and summarize learned rules [4], and on matching against expected patterns by similarities or differences [12] with post analysis, from learned rules [20], a recent highlight is to be extracted. A effort on learning action rules is to split attributes into "hard or soft" [20] or "stable or flexible" [8] to extract actions that may improve the loyalty or profitability of patients. Other work is on action hierarchy [10]. Some other approaches include a combination of two or more methods, for instance, class association rules (or associative classifier) that build classifiers on association rules (A ! C) [6]. External databases are input into characterizing the item sets in class association rules. Clustering is used to reduce the number of learned association rules in cluster association rules. To fit more factors into

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

the KDD process [4], other work is also on, for the transformation from data mining to knowledge discovery [21], and developing a general KDD framework .

III. LITERATURE SURVEY

Many of the current formalizations of data mining algorithms have not reached the goal of facilitating the discovery of concise and interpretable information from large amounts of data, as this is praposed by Charu C. Aggarwal and Co-Authors. One of the reasons for this is that the focus on using purely automated techniques has imposed several constraints on data mining algorithms. For example, any data mining problem such as clustering or association rules requires the speciation of particular problem formulations, objective functions, and parameters. Such systems fail to take the user's needs into account very effectively. This makes it necessary to keep the user in the loop in a way which is both efficient and interpretable. One unique way of achieving this is by leveraging human visual perceptions on intermediate data mining results. Such a system combines the computational power of a computer and the intuitive abilities of a human to provide solutions which cannot be achieved by either.

Even though data mining has been successful in becoming a major component of various business processes as well as in transferring innovations from academic research into the business environment, where the gap between issues that the research group works on and real-world ones is still significant. It is essential for the business and the academic research communities to interact often. KDD-2006 aims at the Data Mining for Business Applications was to gather both researchers and business practitioners and talk about their different perspectives and to share their latest problems and idea. Its expected the result of this is not only to bring them together at KDD but also to create relationships that would continue and grow after the event as well. Domain-driven data mining generally targets actionable knowledge discovery in complex domain issues. It Meta synthesizes its intelligence sources for applicable knowledge discovery. It also identifies challenges and directions for future R&D in the dialogue between academia and business to achieve seamless migration into business world. Knowledge Discovery in Databases (KDD) is a complex interactive process. The promising theoretical framework of inductive databases considers this is essentially the process of querying the data. The query language that can deal either with raw data or patterns that hold data. Patterns that are turns to be the inductive query evaluation process for which constraint-based Data Mining techniques have to be designed.

An inductive query specifies declaratively the desired constraints and algorithms are used to compute the patterns satisfying the constraints in the data. The survey that revealed the important results of this active research domain. Market Surveillance plays important mechanism roles in constructing market models. The existing trading pattern analysis only focuses on data which discloses explicit and high level market dynamics. Between, the existing market surveillance systems available from large exchanges are facing crucial challenges of dynamic and cyber-based misuse, miss-disclosure and misdealing information. Therefore, there is a crucial need to develop innovative and workable methods for smart trading and surveillance. Microstructure pattern analysis studies trading behavior patterns of traders in market microstructure data by utilizing market microstructure knowledge.

The identified market microstructure patterns are then used for powering market trading and surveillance agents for automatically detecting/designing profitable and legal trading strategies or monitoring abnormal market dynamics and trader's behavior. Such trading/surveillance agent-driven market trading/surveillance systems can greatly enhance the analytical, discovery capability of market trading/surveillance than the current predefined rule/alert-based systems. An approach to defining action ability as a measure of interestingness of patterns is proposed. This is based on the concept of an action hierarchy which is defined as a tree of actions with patterns and pattern templates (data mining queries) assigned to its nodes. An applicable patterns method is presented and various techniques for optimizing the discovery process are proposed. Pairs mining targets to mine pair are relationship between entities such as between stocks and markets in financial data mining. It has come up as a kind of promising data mining applications. Due to practical glitches in the real-world pairs mining such as mining high dimensional data and considering user preference, it is challenging to mine pairs of traders in business situations.

This paper presents fuzzy genetic algorithms to deal with these issues. The studies of complex systems have been recognized as one of the greatest challenges for current and future science and technology. As a result, traditional problem-solving methodologies can help deal with them but are far from a mature solution methodology. The theory of qualitativeness-to-quantitativeness meta synthesis has been proposed as a breakthrough and effective methodology for the understanding and problem solving of OCGSs. The concepts of M-Space, M-Computing and M-Interaction which

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

are three key components for studying OCGS and building problem solving algorithms. M Interaction forms are the main problem solving mechanisms of qualitative-to-quantitative meta synthesis; M-Space is the OCGS problem-solving system embedded with M-Interactions, while M-Computing consists of engineering approaches to do analysis, design, and implementation of M-Space, M-Interaction. Most of the literature argues that surprisingness is an inherently , which cannot be measured in objective terms. It is a two fold goal:
(1) To show that it is possible to define objective measures, rather than subjective measures for a discovered rule surprisingness
(2) proposing new ideas and methods for defining objective rule surprisingness measures. 1 Introduction A crucial aspect of data mining is that the discovered knowledge (usually expressed in the form of "if-then" rules) should be somehow interesting, where the term interestingness is arguably related to the properties of surprisingness (unexpectedness), usefulness and novelty of the rule [Fayyad et al. 96]. Traditionally, knowledge action ability has been investigated mainly by developing and improving technical interestingness.
Recently, initial work on technical subjective interestingness and business-oriented profit mining presents general potential, while it is a long-term mission to bridge the gap between technical significance and business expectation. Its proposed that a two-way significance framework for measuring knowledge action ability, which highlights both technical interestingness and domain-specific expectations. Further develop a fuzzy interestingness aggregation mechanism to generate a ranked final pattern set balancing technical and business interests. Real time data mining applications show the proposed knowledge action ability framework can complement technical interestingness while satisfy real user needs.

IV. APPLICABLE KNOWLEDGE DISCOVERY – A SYSTEM OVERVIEW

The data mining process in real time projects is a complex problem solving process. From the stand point of systems and micro economical sub systems, the character of AKD, determines that it is a problem solving methodology, related to optimization, with objectives under given environment.

Let

i) DB be the database of business application.
ii) Ψ be the business problem in database DB.
iii) T be the business problem status of the given problem Ψ
iv) Φ be the problem solution for the given problem Ψ
v) $X=\{x_1,x_2,\ldots\ldots x_l\ldots x_L\}$ be the set of items in DB, where $xl \in X$, for all $1 \le l \le L$
vi) S, be the number of attributes in DB.
vii) $E=\{e_1,e_2,\ldots\ldots e_k,\ldots e_K\}$ be the environment set, where $ek \in E$, for all $1 \le k \le K$, represents a particular environment setting for AKD.
viii) $M=\{m_1,m_2,\ldots\ldots m_n,\ldots m_N\}$ be the data mining method set, where $mn \in M$, for all $1 \le n \le N$
ix) $Pmn=\{p_1{}^{m}{}_{n}, p_2{}^{m}{}_{n},\ldots\ldots p_u m_n ,\ldots p_U{}^{m}{}_{n}\}$ be the identified pattern set, where $pumn \in Pmn$, for all $1 \le u \le U$

Pmn includes discovered patterns in database DB.
pumn includes the pattern discovered by data mining method set mn

x) $\tilde{P}m_n=\{\tilde{p}_1{}^{m}{}_{n}, \tilde{p}_2{}^{m}{}_{n},\ldots\ldots \tilde{p}_u m_n ,\ldots \tilde{p}_U{}^{m}{}_{n}\}$ be an Applicable Pattern Set, where $\tilde{p}_u m_n \in \tilde{P}m_n$, for all $1 \le u \le U$

$\tilde{P}m_n$ includes applicable patterns in database DB.
$\tilde{p}_u m_n$ includes the pattern applicable by data mining method set mn

xi) $P=Pm_1 \cup Pm_2 \ldots\ldots \cup Pm_n$ be the Applicable pattern set

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

xii) R be the data mining problem solving relation from business problem (Ψ) in DB, with problem status T1 to problem solving solutions Φ. Let T2 be the problem status, changed from T1 , over the relation R. This can be represented as follows :

$R : \Psi(T1) \rightarrow \Phi(T2)$ ……..(1)

xiii) Applicable Knowledge Discovery based problem solving process is a state transformation from source data DB to the resulting pattern set P.

The source data can be represented as, from the business problem within the source database DB. That is,

$DB(\Psi \rightarrow DB)$

The resulting pattern set can be represented as, from problem solution set for the business problem to resulting pattern set itself.

That is , $P(\Phi \rightarrow P)$

xiv) Technical Interestingness is represented as t_i that can be applied on given applicable pattern

xv) Techincal Interestingness with threshold is represented as $t_{i,0;}$

xvi) UnsatisfactoryTechnical Interestingness is represented as $\neg t_i$ that can be applied on given applicable pattern

xvii) Business Interestingness is represented as bi that can be applied on given applicable pattern

xviii) Business Interestingness with threshold is represented as bi,0;

xix) Unsatisfactory Business Interestingness is represented as ¬bi that can be applied on given applicable pattern

xx) A be the action, that is taken on arbitrary pattern $p\tilde{}um_n$, which can be represented as $A(p\tilde{}um_n)$

xxi) Int (.) is the interestingness evaluation function

xxii) O(.) is the optimization function to extract those $p\sim \epsilon P\sim$ when Int(p~) can beat a given benchmark.

xxiii) I(.) is the aggregate function that aggregates all elements of a particular aspects of interestingness.

xxiv) Act( ) : is the applicability pattern function for a particular pattern.

Definition 1 (Applicable Patterns) :

Let
$P\tilde{}m_n = \{ p\tilde{}_1{}^m{}_n, p\tilde{}_2{}^m{}_n, \ldots\ldots p\tilde{}_u m_n, \ldots. p\tilde{}_U{}^m{}_n \}$ be an Applicable Pattern Set, where $pu\sim mn \epsilon P\sim mn$, for all $1 \leq u \leq U$, mined by method mn for the given business problem Ψ whose business data comes from DB, in which pu~mn is applicable for the problem solving if it satisfies the following criteria :

1. $t_i(p_u\tilde{}) \geq t_{i,0}$; "≥" indicates the pattern $p_u\tilde{}$ can beat technical interestingness ti with threshold ti,0;

2. $b_i(p_u\tilde{}) \geq b_{i,0}$; "≥" indicates the pattern $p_u\tilde{}$ can beat business interestingness bi with threshold bi,0;

3. From (1) $R : \Psi(T1) \rightarrow \Phi(T2)$

   Or this can be written as $R : T1 \rightarrow T2$

$$R : T1 \xrightarrow{A(p\sim umn)} T2$$

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

That is, the patter can support business problem solving by taking action A, and correspondingly, transform the problem status from the initially non optimal status T1 to the greatly improved T2

Definition 2 (Applicable Knowledge Discovery) :

(i) Applicable Knowledge Discovery is an optimization process that eventually leads to applicable pattern set P~ in iterative method on consideration of given business environment and related problem states

$$AKD\ e, T, m \in M : DB \xrightarrow{e, T, m_n} \begin{cases} Pm_n \\ Op \epsilon P\ e, T, m \in M \quad (Int\ (p)) \\ P\sim \end{cases} \quad \ldots\ldots\ldots \ (2)$$

where P=Pm1 U Pm2 ……. U Pmn be the Applicable pattern set

Int (.) is the interestingness evaluation function

O(.) is the optimization function to extract those $p\sim \epsilon\ P\sim$ when Int(p~) can beat a given benchmark.

(ii) For a given pattern Int(p) can be further measured in terms of
a) technical interestingness (t_i (p))
b) business interestingness (b_i (p))

This can be written as $Int(p)=I(\ t_i\ (p), b_i\ (p)\ )$ ……..……. (3)

Here I is aggregating interestingness aspects of pattern p , that is both technical and business interestingness elements t_i (p), b_i (p)

(iii) Int(p) can be described in terms of the following patters, from both technical (t) and business (b) perspectives
a) objective (o)
b) subjective (s)

Therefore, Int (p) = $I\ (t_o(p), t_s(p), b_o(p), b_s(p))$

$\xrightarrow{}\ t_o(x,p) \wedge t_s(x,p) \wedge b_o(x,p) \wedge b_s(x,p)$ ……..……. (4)

Here, $t_o()$ is objective technical interestingness
$t_s()$ is subjective technical interestingness
$t_b()$ is objective business interestingness
$t_b()$ is subjective business interestingness
$I \xrightarrow{}\ \wedge$, represents the aggregation.

For practical applications, to() , ts() , tb() and tb() are regarded as independent of each other. Consider their normalization is expressed by ^.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

iv) So we will get the following normalized form with different coefficients for each function (called weights) for the whole interestingness of pattern P can be represented as follows :

$$\text{Int}(p) \longrightarrow I^{\wedge}(t_o^{\wedge}(p), t_s^{\wedge}(p), b_o^{\wedge}(p), b_s^{\wedge}(p))$$

$$= \alpha t_o^{\wedge}(p) + \beta t_s^{\wedge}(p) + \gamma b_o^{\wedge}(p) + \delta b_s^{\wedge}(p) \quad \text{.....} \quad (5)$$

Here α, β, γ, δ are weights for interestingness's to() , ts() , tb() and tb() respectively.

v) So, the Applicable Knowledge Discovery optimization problem can be re-written as follows :

from equation (2):

$$\text{AKD}\ e, T, m \in M : DB \xrightarrow{e, T, mn} Pmn \longrightarrow Op \epsilon P\ e, T, m \in M\ (\text{Int}(p)) \longrightarrow P\sim \quad \text{.....} \quad (2)$$

That is,

$$\text{AKD}\ e, T, m \in M \longrightarrow Op \epsilon P\ e, T, m \in M\ (\text{Int}(p))$$

$$\longrightarrow Op \epsilon P\ (\alpha t_o^{\wedge}(p) + \beta t_s^{\wedge}(p) + \gamma b_o^{\wedge}(p) + \delta b_s^{\wedge}(p)) \quad (\text{from } 5)$$

This can be generalized to all patters as follows

$$\text{AKD}\ e, T, m \in M \longrightarrow Op \epsilon P\ e, T, m \in M\ (\text{Int}(p))$$

$$\longrightarrow O\ (\alpha to^{\wedge}(\ ) \wedge \beta t_s^{\wedge}(\ ) \wedge \gamma b_o^{\wedge}(\ ) \wedge \delta b_s^{\wedge}(\ )) \quad \text{.....} \quad (6)$$

Definition 3 (Applicability of a pattern)

(i) For a given pattern p, the applicability is measured by act (p).

$$act(p) = Op \epsilon P\ (\text{Int}(p))$$

$$\longrightarrow O\ (\alpha t_o^{\wedge}(p) \wedge \beta t_s^{\wedge}(p) \wedge \gamma b_o^{\wedge}(p) \wedge \delta bs^{\wedge}(p)) \quad (\text{from } 6)$$

$$\longrightarrow O\ (\alpha t_o^{\wedge}(p)) \wedge O\ (\beta t_s^{\wedge}(p)) \wedge O(\gamma b_o^{\wedge}(p)) \wedge O\ (\delta b_s^{\wedge}(p))$$

$$\longrightarrow t_o^{act} \wedge t_s^{act} \wedge b_o^{act} \wedge b_s^{act}$$

$$\longrightarrow t_i^{act} \wedge b_i^{act} \quad \text{.....} \quad (7)$$

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Here t_o^{act} , t_s^{act} , b_o^{act} and b_s^{act} measure the respective applicable performance.

(ii) Factors like satisfaction and Business performance are considered in applicable frequent trading pattern mining.

The satisfaction factor further viewed as support and confidence.
And business performance like sharpe ratio.

Assume that they are independent. Then, by using maximal manner, we can expect an applicable trading pattern to concurrently satisfy these metrics.

There is minimal amount of inconsistency that can be expected in different aspects, often find identified patterns fitting in one of the following subsets :

From equation ❼ applicable pattern is general for a particular Interestingness.

$$Act(p) \longrightarrow t_i^{act} \wedge b_i^{act} \qquad \ldots\ldots\ldots\ldots\ldots\ldots ❼$$

If the interestingness for the above applicable pattern is Int (p), then this can be written as

$$Int\ (p) \longrightarrow \{\ t_i^{act} , b_i^{act}\ \} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots ❽$$

the element of unsatisfactory is introduced in the interestingness, then the above pattern will take following combination of sub-patterns If

$$Int\ (p) \longrightarrow \{\{t_i^{act} , b_i^{act}\} ,\{\neg t_i^{act} , b_i^{act}\} , \{ t_i^{act} , \neg b_i^{act}\} ,\{\neg t_i^{act} , \neg b_i^{act}\}\} \quad \ldots.❾$$

Here ¬ represents unsatisfactory.

It's a challenge in real world mining that the existence of most applicable patterns that are associated with "optimal" t_i^{act} and b_i^{act} . It's very general that a pattern with considerable $t_i(\ )$ is associated with unconfident $b_i(\ )$. On the contrary to this, the patterns with low $t_i(\ )$ are associated with confident $b_i(\ )$**.** By taking all combinations of confident and unconfident of patterns, AKD targets patterns confirming the relationship $\{t_i^{act} , b_i^{act}\}$. This indicates the necessity of dealing intelligently with possible conflict and uncertainty among respective interestingness elements. Though it's necessary to involve SMEs in domain knowledge and Domain Experts to tune response thresholds and draw balance differences between $t_i(\ )$ and $b_i(\ )$. The development of various techniques to balance and combine all types of interestingness metrics to generate uniform, balanced, and interpretable mechanisms for measuring knowledge deliverability, extracting and selecting resulting patterns, is yet another issue.

There is a clear urgency to develop an interestingness aggregation methods, namely the I-function (or " ∧ ") to aggregate all elements of interestingness. Each of the interestingness categories may be instantiated into more than one metric. Here the aggregation does not mean the essential combination into a single super measure, rather indicating the satisfaction of all respective components during the AKD process if possible. There could be various methods of doing the aggregation, for example, methods like business-export-based voting, or more quantitative methods like multi objective optimization methods. Knowledge applicability also needs to cater for the semantic aspect of the identified applicable patterns, besides the measurement. The Business Rule Model [23] defined in OWL-S [1] defines the conversation from an identified pattern to a business rule .

To describe a business rule, the following aspects are necessary to notice:

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

a) Object : On what object (s), the actions are performed, along with predicates sothat the range can be limited;
b) Condition : Situations under which the actions can be taken on the objects, with predicates to specify the conditions;
c) Operation : What actions are to be taken on the objects, with predicates to deliver the specific decision-making activities.

Applicable patterns are converted into business rules, after following specification, as a form of deliverable, which not only enhances interpretation but also indicates what actions can be taken on what objects under what conditions.

/Business Rule Specification */
< business_rule > :: = < object > $^{+}$ < condition > * < operation > $^{+}$

< object > :: = (All | Any | Given | …….)
< condition > :: = (Satisfy | Related | and | …….)
< operation > :: = (Alert | Action | …….)

V. APPLICABLE DATA MINING RULE DISCOVERY

A. Discovery of applicable patterns

The following steps describe the discovery process of applicable patterns by using action trees

i) Building an Action Tree

For a given application an action tree must be built and maintained later on. If the hierarchical approach to action specification is considered, then it is recommended to maintain a hierarchy of actions from more general actions at the top of the hierarchy to more specific actions at the bottom.

For example consider a customer purchase data for a supermarket application and the actions that the store manager can take based on this data. All the possible actions that a supermarket manager can take are grouped into the customer related actions , product stocking actions, advertising actions, promotion related actions...etc. This broad classification of actions can be further subdivided into more specific actions, known as sub actions. A fragment of such a hierarchy for the supermarket application is presented in Figure 1.

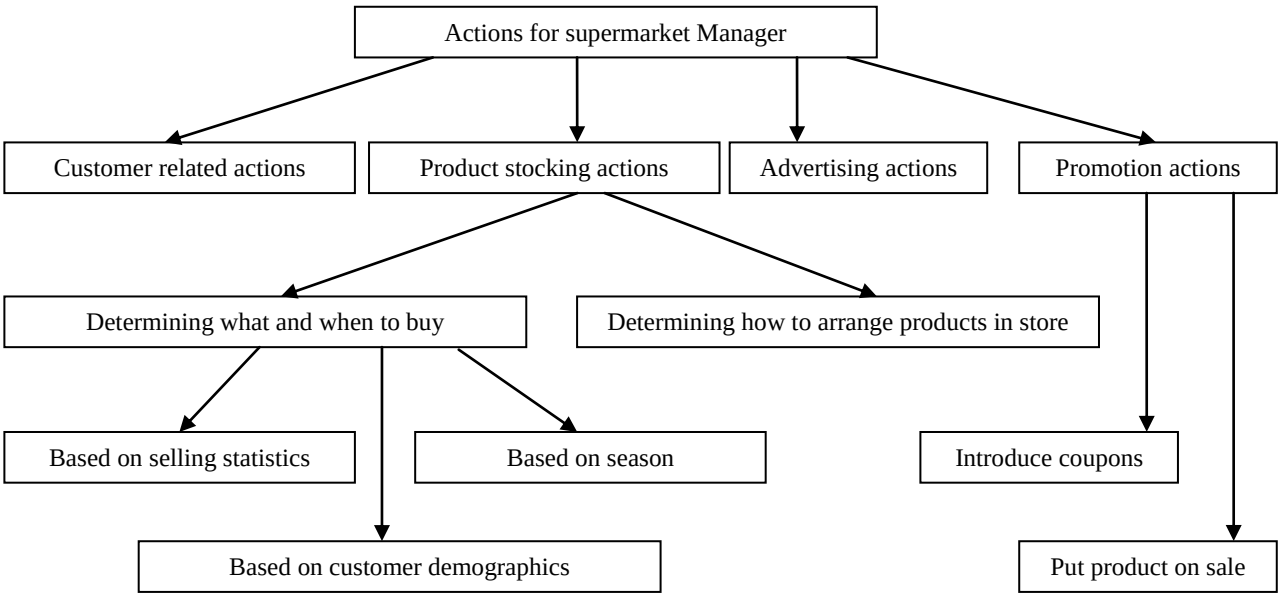


Figure 1 : Fragment of an action tree for the supermarket management.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

To demonstrate the process of the top-down construction of an action hierarchy, consider the node "product stocking actions". As above figure dipicts, the action is divided into two subcategories ,as mentioned below

a) Determining what and when to buy
b) Determining how to arrange products in the store

The first sub category is sub divided further into sub actions like

a) Based on selling statistics
b) Based on season
c) Based on customer demographics

ii) Applicable Attributes

Specifying the applicable patterns using action trees, can be achieved in two ways: 1) Assign individual patterns to various nodes of the tree, thus declaring these patterns to be applicable. Patterns have to be specified in some pattern description language.

For example, The following association rule specifying the extent to which families with small children buy sweets, to the action node in Figure 1 "(Determining what and when to buy) Based on customer demographics" :

ChildrenAgeLess 6 => CategorySweets (0.55,0.01) ……❿

Consider the request "Find all rules in customer prucahsae **data** specifying which product categories the customers with children of different ages are buying". This can be expressed in the pattern description language as

ChildrenAge * => Category (0.5, 0.01) ……..……..⓫

iii) Assigning Data Mining Queries

The nodes of Action Tree, are assigned with the data mining queries defining applicable patterns for the specific Figure 1, could be the query ⓫. Additional examples of data mining queries expressed in pattern template language are :actions. A possible data mining query assigned to the node "Based on customer demographics" of the tree in

a) Query : "Find what kind of product categories sell well on different days of week" (assigned to the action "Based on season") :

DayOfWeek => Category + (0.4, 0.01) …….....…….⓬

b) Query : "Find 'Cross-selling' categories, that are, find categories of products that are selling together" (assigned to the action "Determining how to arrange products in the store"):

Category+ => Category+ (0.5, 0.01) …….....……⓭

iv) Executing data mining queries :

The pattern discovery process in a given attributed tree action, consists of the traversal of the whole action tree, using depth-first search and execution of all the data mining queries. The discovered applicable patterns are written to the files associated with data mining queries.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

B. **Discovery optimization**

Whenever the data changes in the database happens, then re-executing all the data mining queries, is computationally expensive task. This is true for big applications with large action trees and many data mining queries. The following two optimization techniques are going address this issue.

i) **Partial tree traversal:**

The partial traversal of an action tree, is the natural optimization of the action tree traversal technique. In this method, only ht nodes of the tree selected by the user are traversed and only those data mining queries that are assigned to these nodes are executed. These nodes can be selected as individual nodes or as belonging to the ser specified sub tree. The applications, in which there is no need to keep patterns up to date all the time, can be used partial tree traversal approach extensively. So, the data mining queries can be executed whenever there is a need to consider some specific action, only then the data mining queries assigned to that action must be re-executed to supply the user with the latest patterns to help make decisions. (On Demand basis execution).

ii) **Triggers:**

The data mining queries should be re-execute only when "substantial" changes occur in the data that affect the patterns discovered by the queries , especially in the field of stock market analysis applications. This would save resources by avoiding unnecessary executions of the queries not affected by data related changes. One of the ways to detect such changes is to use the data monitoring method, which uses extended triggers (also known as $DMDT^2$ triggers)

$DMDT^2$ triggers are defined as follows :

Let

a) D be the data stored in the data
b) ΔD be the new data to be added to this database D

In supermarket application, D could be the supermarket customer purchase data for the last 6 months, and ΔD could be the daily customer purchase data which is recorded daily in the central database.

An extended trigger could be of following form:

WHEN new data ΔD becomes available
IF "Significant changes" in data, are found when ΔD is added to the old data D
THEN execute the data mining query

These triggers are called "extended" because they are extensions of classical triggers used in the active databases.

C. **MSCM-ADM**

When an application has multiple sub systems and heterogeneous data sources from where the data will be input to the application, face possible issues as follows :

a) Integration issues while merging sub systems to the main one and inter related sub systems themselves
b) Costly because of possible data loss and integration methods
c) Synchronization issues while integrating sub systems
d) Scanning issues because of large volumes of data.

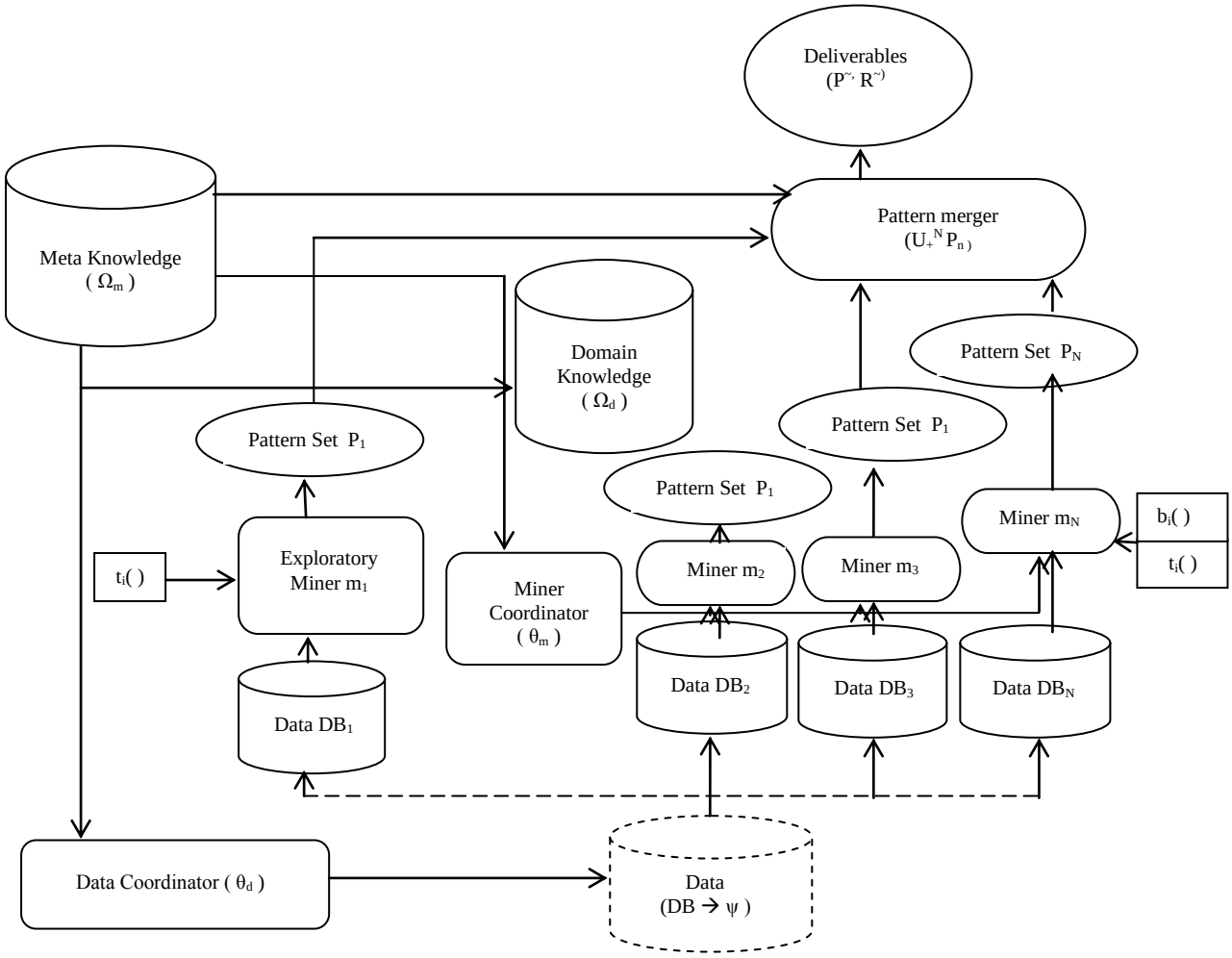
ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

e) Scanning issues of the dataset , at the module level.
To address the above issues, the proposed framework is Multi Source + Combined Mining based Applicable Data Mining (MSCM-ADM) Rule. The following figure depicts a overview of MSCM-ADM Rule.



1) MSCM-ADM rule discovers applicable knowledge either in multiple data sets or data sub sets (DB1, DB2, ... DB_N) through partitions.

2) For Mining Exploration (m1) process, one of the data sets or certain partial data , by random DBn , where $1 \leq n \leq N$ is selected based on Domain Knowledge, Business Understanding and Goal Definition.

3) The exploration results are used to guide either data partition or data set management through a Data Coordinator Agent ( ) (coordinating data partition and / or data set /feature selection in terms of iterative mining processes) (24) and to design strategies for managing and conducting "Parallel Pattern Mining" on each data set or subset and /or "Combined Mining" [10] on relevant remaining data sets.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

4) The deployment of methods m_n, which could be either in parallel or combined, is determined by data / business understanding and objectives.

5) After the mining of all data sets, patterns P_n identified from individual data sets are merged and extracted into final deliverable ($P^{\sim}$, $R^{\sim}$)

6) MSCM-ADM can be expressed as follows :

MSCM-ADM :

$$\underbrace{DBn\ [\ DB \xrightarrow{\Theta} DBn\ ]}_{N} \xrightarrow{e, t_{i,n}(\ )[u_{i,n}(\ )], m_n, \Omega_m \{ Pn\}} \xrightarrow{e, b_{i,n}(\ ), U^{N} P_n, \Omega_d, \Omega_m} P^{\sim}, R^{\sim} \ldots\ldots (14)$$

Where

a) $t_{i,n}$ and $b_{i,n}$ are technical and business interestingness of model m_n on data set or sub set n,
b) [$i_{i.n}$ ( )] indicates the alternative checking of unified interestingness as in Unified Interestingness based Applicable Data Mining rule.
c) $U^{+N} P_n$ is the merger function
d) Θ indicates the data partition if the source data needs to be split.

The MSCM based ADM (MSCM-ADM) - Algorithm :

INPUT :

a) Target data set DB
b) Business Problem ψ
c) Thresholds (to,0 , ts,0 , bo,0 and bs,0)

OUTPUT :

a) Applicable patterns $P^{\sim}$
b) Business Rules $R^{\sim}$

Algorithm steps :

Step 1) Partition whole source data into N data sets DB_n where $1 \leq n \leq N$

Step 2) **Data Set-n Mining :**
Extracting general patterns P_n on data set or sub set DB_n
FOR l = n to N
a) Develop modelling method m_n with technical interestingness $t_{i,n}$ ( ) ;
That is , t_o( ), t_b( ) or unified $i_{i,n}$ ( )
b) Employ method m_n on the environment e and data DB_n engaging meta-knowledge Ω_m ;
c) Extract the general pattern set P_n;
END FOR

Step 3) **Pattern Merger:**
Extracting applicable patterns $P^{\sim}$
FOR l = n to N

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

a) Design the pattern merger function $U^{+N}P_n$ to merge all patterns into P by involving domain and meta knowledge Ω_d and Ω_m, and business interestingness $b_i(\ )$;
b) Employ the method $U^{+}P_n$ on the pattern set P_n;
c) Extract the applicable pattern set $P^{\sim}$;

END FOR

Step 4) **Converting patterns $P^{\sim}$ to business rules $R^{\sim}$:**

The MSCM-ADM framework can also be instantiated into many mutations. For example, for a given large volume of data, SMCD-ADM can be instantiated into (Data Partition) + (Unsupervised) + (Supervised-based ADM) by integrating data partition into combined mining.

Example :

a) The whole data set is partitioned into several data subsets based on the data or business understanding and domain knowledge jointly by data miners and domain exports, say data sets 1 and 2.

b) An (unsupervised learning) method is used to mine one of the preference data sets, day data set 1. Some of the mined results are then used to design new variables for processing the other data set.

c) (Supervised Learning) is further conducted on data set 2 to generate applicable patterns by checking both technical and business interestingness.

d) The individual patterns mined from both data subsets are combined into deliverables.

VI. RESULTS & DISCUSSIONS

The randomly generated medical data is been tested by MSCM-ADM method. 55,800 patients with their demographic are present in cleaned sample data and 711 traditional associations mined. Combined associations cannot be discovered by traditional association rule techniques. When compared with the single associations from respective data sets, the combined associations and clusters are much more workable than single rules presented in the traditional way. They have vital information from multiple aspects rather than from a single one, or a collection of separated single rules.

For example, the following combined association shows that patients aged 65 or more, whose arrangement method is of “Smoking” plus “regular”, then they have more chances of getting cancer. This can be classified as “C” (High Risk of life). Clearly this pattern combines heterogeneous information regarding the specific group of the patient’s demographic method.

{ x= age ; 65 +, y = Smoking & Repeated + Cancer → c = C }

Combined patterns can be transformed into operable business rules that may indicate direct actions for business decision making. For example, for above combined association, it connects key business elements with segmented patient characteristics, and we generate following business rule by extending the Business Rule specification:

Delivering Business Rules:

Patient Demographic – Combination business rules

Assumptions :

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

I be the set of number of valid patients
Patients in set I are having habit of smoking , either regularly or occasionally.

Algorithm :

FOR ALL patients i (i ϵ I is the number of valid patients)

Condition :

He / she under arrangement of "Smoking" and "regular"
And
He / she is also having "Cancer"

Operation :

Alert = "He / She has 'High' risk of life in short timeframe"
Action = "Try to avoid smoking habit or take medical advise"

END ALL

The converted business rules are deliverables presented to business people. They are convenient and it is easy for clients to embed them into their routine business processes and operational systems for filtering patients and monitoring the cancer patients.

Our clients feel comfortable in understanding, interpreting, and actioning these business rules than those patterns directly mined in the data. Thus combined patterns are more business friendly and indicate much more straightforward decision-making actions to be taken by business analysts in the business world, while this cannot be achieved by patterns identified by traditional methods.

In addition, use of combined mining leads to combined patterns that are consisting of various attributes from different business units or by partitioning into organized segments. Through attribute segmentation, it is manageable to differentiate attribute impact on business objectives, informative patterns and more operable decision-making actions.

VII. CONCLUSION

In this paper the Applicable Data Mining concepts, processes, applicability of patterns, and operable deliverables are clearly defined and with these components, we have proposed MSCM ADM framework capable of handling various business problems and applications. This framework supports closed-optimization-based problem solving from a business problem or environment definition to applicable pattern discovery to operable business rule conversion. Deliverables extracted in this way are not only of technical significance but also are capable of smoothly integrating into business processes. This framework is general, flexible, and workable to be instantiated into various approaches for tackling complex data and business applications. Substantial experiments in significant data mining applications such as financial data mining and mining social security data have shown that the proposed framework have the potential to handle the limitations in existing methodologies and approaches. Following the D^3M theory, there are many issues to be studied, for instance, defining operable business rules by involving ontological techniques for representing both syntactic and semantic components.

REFERENCES

[1] K. Breitman, M. Casanova, and W. Truszkowski, "Semantic Web.Springer", 2007.
[2] J.F. Boulicaut and B. Jeudy, "Constraint-Based Data Mining," The Data Mining and Knowledge Discovery Handbook, pp. 399-416, Springer, 2005.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

[3] Q. Yang, J. Yin, C. Ling, and R. Pan, "Extracting Actionable Knowledge from Decision Trees," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 43-56, Jan. 2007.
[4] B. Liu and W. Hsu, "Post-Analysis of Learned Rules," Proc. Nat'l Conf. Artificial Intelligence/Innovative Applications of Artificial Intelligence Conf. (AAAI/IAAI), 1996.
[5] C. Aggarwal, "Towards Effective and Interpretable Data Mining by Visual Interaction," ACM SIGKDD Explorations Newsletter, vol. 3, no. 2, pp. 11-22, 2002.
[6] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," Data Mining and Knowledge Discovery, vol. 2, no. 4, pp. 311-324, 1998.
[7] R. Hilderman and H. Hamilton, "Applying Objective Interestingness Measures in Data Mining Systems," Proc. Symp. Principles of Data Mining and Knowledge Discovery (PKDD), pp. 432-439, 2000.
[8] B. Lent, A.N. Swami, and J. Widom, "Clustering Association Rules," Proc. 13th Int'l Conf. Data Eng., pp. 220-231, 1997.
[9] L. Cao, "Domain-Driven Actionable Knowledge Discovery," IEEE Intelligent Systems, vol. 22, no. 4, pp. 78-89, July/Aug. 2007.
[10] G. Adomavicius and A. Tuzhilin, "Discovery of Actionable Patterns in Databases: The Action Hierarchy Approach," Proc.Int'l Conf. Knowledge Discovery and Data ining (KDD '97), pp. 111-114, 1997.
[11] L. Cao, "Domain-Driven Data Mining: Empowering Actionable Knowledge Delivery," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '09) Tutorial, 2009.
[12] B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing Subjective Interestingness of Association Rules," IEEE Intelligent Systems,vol. 15, no. 5, pp. 47-55, Sept./Oct. 2000.
[13] L. Cao and Y. Ou, "Market Microstructure Pattern Analysis for Powering Trading and Surveillance Agents," J. Universal Computer Science, vol. 14, no. 14, pp. 2288-2308, 2008.
[14] L. Cao, P. Yu, C. Zhang, and H. Zhang, Data Mining for Business Applications. Springer, 2008.
[15] H. Kargupta, B. Park, D. Hershbereger, and E. Johnson, "Collective Data Mining: A New Perspective toward Distributed Data Mining," Advances in Distributed Data Mining, H. Kargupta and P. Chan, eds., AAAI/MIT Press, 1999.
[16] E. Omiecinski, "Alternative Interest Measures for Mining Associations," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 1, pp. 57- 69, Jan./Feb. 2003.
[17] O.G. Ali and W. Wallace, "Bridging the Gap between Business Objectives and Parameters of Data Mining Algorithms," Decision Support Systems, vol. 21, pp. 3-15, 1997.
[18] P. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns," Proc. ACM SIGKDD, pp. 32-41, 2002.
[19] A. Freitas, "On Objective Measures of Rule Surprisingness," Proc. Second European Symp. Principles of Data Mining and Knowledge Discovery (PKDD '98), pp. 1-9, 1998.
[20] B. Liu, W. Hsu, and Y. Ma, "Pruning and Summarizing the Discovered Associations," Proc. ACM SIGKDD, 1999.
[21] L. Cao, Y. Zhao, C. Zhang, and H. Zhang, "Activity Mining: From Activities to Actions," Int'l J. Information Technology and Decision Making, vol. 7, no. 2, pp. 259-273, 2008. [12] L. Cao, P. Yu, C. Zhang, and Y. Zhao, Domain Driven Data Mining. Springer, 2009.
[22] M. Ankerst, "Report on the SIGKDD-2002 Panel the Perfect Data Mining Tool: Interactive or Automated?" ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 110-111, 2002.
[23] L. Cao, Y. Zhao, and C. Zhang, "Mining Impact-Targeted Activity Patterns in Imbalanced Data," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 8, pp. 1053-1066, Aug. 2008.
[24] Longbing Cao, Ana L.C. Bazzan, Vladimir Gorodetsky, "Agent and Data Mining Interaction", 6th International Workshop on Agents and Data Mining Interaction, ADMI 2010, Toronto, ON, Canada, May 2010, Revised Selected Papers, 2010