Actionable Knowledge Discovery

Longbing Cao
University of Technology Sydney, Australia

INTRODUCTION

Actionable knowledge discovery is selected as one of the greatest challenges (Ankerst, 2002; Fayyad, Shapiro, & Uthurusamy, 2003) of next-generation knowledge discovery in database (KDD) studies (Han & Kamber, 2006). In the existing data mining, often mined patterns are nonactionable to real user needs. To enhance knowledge actionability, domain-related social intelligence is substantially essential (Cao et al., 2006b). The involvement of domain-related social intelligence into data mining leads to *domain-driven data mining* (Cao & Zhang, 2006a, 2007a), which complements traditional data-centered mining methodology. Domain-related social intelligence consists of intelligence of human, domain, environment, society and cyberspace, which complements data intelligence. The extension of KDD toward domain-driven data mining involves many challenging but promising research and development issues in KDD. Studies in regard to these issues may promote *the paradigm shift of KDD from data-centered interesting pattern mining to domain-driven actionable knowledge discovery*, and *the deployment shift from simulated data set-based to real-life data and business environment-oriented* as widely predicted.

BACKGROUND

In the last decades, data mining, or KDD, has become a prominent, exciting research and development area in the field of information technology. Data-centered data mining has experienced rapid development in various aspects such as data mined, knowledge discovered, techniques developed, and applications involved. Table 1 illustrates such key research and development progress in KDD.

A typical feature of data-centered data mining is that KDD is presumed to be an automated process of identifying interesting hidden patterns in public data sets. It targets the production of automatic algorithms as well as methods that extract patterns of certain technical significance. As a result, algorithms and the tools developed lack the capability to adapt to real-life environmental constraints and dynamics. Thousands of patterns and algorithms have been published in academia, but unfortunately very few of them have been transferred into real business use.

Increasing numbers of KDD researchers and developers have realized the limitation of traditional data mining methodologies (Ankerst, 2002; Fayyad et al., 2003), noted the gap between business and academic interests (Gurali & Wallace, 1997). The research on the challenges of KDD,

Table 1. An overview of data mining

Dimension	Key research progress
Data mined	• Relational, transactional, object-relational, active, temporal, spatial, time-series, heterogeneous, legacy, Web, and so forth. • Stream, spatiotemporal, multimedia, ontology, event, activity, link, graph, text, sensor, and so forth.
Techniques studied	• Database, machine learning, or statistics-oriented, say Neural Network, Bayesian network, Support Vector Machine, Rough Set, and so forth. • Association, frequent pattern analysis, multidimensional and OLAP analysis methods, classification, cluster analysis, outlier detection, visualization, and so forth. • Scalable data mining, stream data mining, spatiotemporal data mining, multimedia data mining, biological data mining, text and Web mining, privacy-preserving data mining, event mining, link mining, ontology mining, granule mining, and so forth.
Knowledge discovered	• Characters, associations, classes, clusters, discrimination, trends, deviation, outliers, exceptions and so forth.
Application involved	• Engineering, retail market, telecommunication, banking, fraud detection, intrusion detection, stock market, social security, bio-informatics, defense, Web services, biological, social network analysis, intelligence and security, and so forth. • Enterprise data mining, cross-organization mining, online mining, dynamic mining, and so forth.

plus trustworthy and workable KDD methodologies and techniques have therefore become a significant and productive direction of KDD research. In the panel discussions of SIGKDD 2002 and 2003 (Ankerst, 2002; Fayyad et al., 2003), a couple of important challenges for extant and future data mining were identified. Among them, actionable knowledge discovery is viewed as one of the key foci, because it not only provides an important tool to business decision makers for performing appropriate actions, but also delivers reliable and actionable outcomes to businesses. However, it is not a simple task to extract actionable knowledge utilizing traditional KDD methodologies. This situation results partly from the assumption that traditional data mining is a data-centered trial-and-error process (Ankerst, 2002), data and technical interestingness have been taken as two of the major targets and criteria in algorithm and pattern development.

To bridge the gap between business and academia, it is important to understand the difference between objectives and result evaluation of data mining in research and real-world applications. In the business world, KDD must answer the question "What Makes Knowledge Identified Interesting to Businesses" (Silberschatz & Tuzhilin, 1996) not only from the technical angle but also from a business perspective. Real-world data mining needs to identify patterns in constrained environments and satisfy not only technical significance (Freitas, 1998; Hilderman & Hamilton, 2000; Omiecinski, 2003; Padmanabhan & Tuzhilin, 1998), but business expectations (Ghani & Soares, 2006; Cao & Zhang, 2007a; Kleinberg, Papadimitriou, & Raghavan, 1998). The difference referred to above is exemplified through several key aspects (Cao & Zhang 2007a), for example, KDD problem, context, patterns, mining processes, objective and subjective interestingness, business expection, balancing multiple objectives, and infrastructure supporting business-oriented mining.

To deal with this difference, experience and lessons learned in real-world data mining (Cao & Zhang 2006a, 2007a, 2007c, 2008a, 2008b) show the significance of involving domain-specific data, domain, human and cyberspace intelligence (Cao et al., 2006b). For instance, domain-related social intelligence may consist of the involvement of domain knowledge (Yoon, Henscen, Park, Makki, 1999) and experts (Cao & Dai, 2003; Han & Kamber, 2006), the consideration of constraints (Boulicaut & Jeudy, 2005), and the development of in-depth patterns (Lin & Cao, 2006). It is essential to filter subtle concerns while capturing incisive issues. Through the thorough scrutiny of domain-specific intelligence and correct involvement of domain-specific intelligence into KDD, a streamlined data mining methodology emerges to discover the hidden core of a problem. These form the grounds of *domain-driven data mining* for next-generation KDD.

KNOWLEDGE ACTIONABILITY

In order to reflect business concerns, a *two-way significance framework* (Luo, Cao, Ni, & Liu, 2007) is proposed for measuring knowledge actionability. The two-way significance framework presents a straightforward nevertheless important definition of knowledge actionability. This means highlighting the involvement of business expectations (Tzacheva & Ras, 2005; Wang, Zhou, & Han, 2002) into traditional technical-only significance scheme (Yang, Yin, Lin, & Chen, 2003). In addition, the two-way significance is reflected in terms of both objective (Freitas, 1998; Hilderman & Hamilton, 2000), subjective (Liu, Hsu, Chen, & Ma, 2000), and multi-objective (Freitas, 2004; Tuzhilin, 2002) perspectives. As a result, actionable knowledge identified is not only based on a solid technical foundation, for instance, of recognized statistical significance, but enables business users to take appropriate actions which will be to their advantage.

Definition 1. (*Knowledge Actionability*) *Let x be an itemset in dataset X, given a mined pattern p associated with x, actionable capability x.act(p) is described as the satisfaction of both technical interestingness x.tech_int(p) and business expectation x.biz_int(p).*

$$\forall x \in X, \exists p : x.tech\_int(p) \wedge x.biz\_int(p) \rightarrow x.act(p) \quad (1)$$

Further, knowledge actionability is instantiated in terms of objective (*_obj*()) and subjective (*_sub*()) perspectives from both technical and business sides:

$$\forall x \in X, \exists p: x.tech\_obj(p) \wedge x.tech\_subj(p) \wedge biz\_obj(p) \wedge biz\_sub(p) \rightarrow x.act(p) \quad (2)$$

However, it is not rare to discover that incompatibility and uncertainty exist in bridging the gap between business and academia in real-world data mining. To solve these issues, we propose (Luo, et al., 2007) fuzzy aggregation of business expectation and technical significance, generating a fuzzy ranking of patterns reflecting and balancing both technical and business concerns. As demonstrated in mining actionable trading patterns in market order book data (Lin & Cao, 2006), this approach presents promising options for resolving the issue of business needs, as well as bridging the gap between the business and technical side.

KEY COMPONENTS

Based on our experience and lessons learned in developing domain-driven data mining for real-world applications, including the discovery of actionable trading patterns in stock markets (Lin & Cao, 2006) and activity patterns