# Chapter 1
# Introduction to Domain Driven Data Mining

Longbing Cao

**Abstract** The mainstream data mining faces critical challenges and lacks of soft power in solving real-world complex problems when deployed. Following the paradigm shift from 'data mining' to 'knowledge discovery', we believe much more thorough efforts are essential for promoting the wide acceptance and employment of knowledge discovery in real-world smart decision making. To this end, we expect a new paradigm shift from 'data-centered knowledge discovery' to 'domain-driven actionable knowledge discovery'. In the domain-driven actionable knowledge discovery, ubiquitous intelligence must be involved and meta-synthesized into the mining process, and an actionable knowledge discovery-based problem-solving system is formed as the space for data mining. This is the motivation and aim of developing *Domain Driven Data Mining* ($D^3M$ for short). This chapter briefs the main reasons, ideas and open issues in $D^3M$.

## 1.1 Why Domain Driven Data Mining

Data mining and knowledge discovery (data mining or KDD for short) [9] has emerged to be one of the most vivacious areas in information technology in the last decade. It has boosted a major academic and industrial campaign crossing many traditional areas such as machine learning, database, statistics, as well as emergent disciplines, for example, bioinformatics. As a result, KDD has published thousands of algorithms and methods, as widely seen in regular conferences and workshops crossing international, regional and national levels.

Compared with the booming fact in academia, data mining applications in the real world has not been as active, vivacious and charming as that of academic research. This can be easily found from the extremely imbalanced numbers of pub-

Longbing Cao
School of Software, University of Technology Sydney, Australia, e-mail: lbcao@it.uts.edu.au

lished algorithms versus those really workable in the business environment. That is to say, there is a big gap between academic objectives and business goals, and between academic outputs and business expectations. However, this runs in the opposite direction of KDD's original intention and its nature. It is also against the value of KDD as a discipline, which generates the power of enabling smart businesses and developing business intelligence for smart decisions in production and living environment.

If we scrutinize the reasons of the existing gaps, we probably can point out many things. For instance, academic researchers do not really know the needs of business people, and are not familiar with the business environment. With many years of development of this promising scientific field, it is time and worthwhile to review the major issues blocking the step of KDD into business use widely.

While after the origin of *data mining*, researchers with strong industrial engagement realized the need from 'data mining' to 'knowledge discovery' [1, 7, 8] to deliver useful knowledge for the business decision-making . Many researchers, in particular early career researchers in KDD, are still only or mainly focusing on 'data mining', namely mining for patterns in data. The main reason for such a dominant situation, either explicitly or implicitly, is on its originally narrow focus and overemphasized by innovative algorithm-driven research (unfortunately we are not at the stage of holding as many effective algorithms as we need in the real world applications).

Knowledge discovery is further expected to migrate into *actionable knowledge discovery* (AKD) . AKD targets knowledge that can be delivered in the form of business-friendly and decision-making actions, and can be taken over by business people seamlessly. However, AKD is still a big challenge to the current KDD research and development. Reasons surrounding the challenge of AKD include many critical aspects on both macro-level and micro-level.

On the macro-level, issues are related to methodological and fundamental aspects, for instance,

- An intrinsic difference existing in academic thinking and business deliverable expectation; for example, researchers usually are interested in innovative pattern types, while practitioners care about getting a problem solved;
- The paradigm of KDD, whether as a hidden pattern mining process centered by data, or an AKD-based problem-solving system ; the latter emphasizes not only innovation but also impact of KDD deliverables.

The micro-level issues are more related to technical and engineering aspects, for instance,

- If KDD is an AKD-based problem-solving system, we then need to care about many issues such as system dynamics, system environment, and interaction in a system;
- If AKD is the target, we then have to cater for real-world aspects such as business processes, organizational factors, and constraints.

In scrutinizing both macro-level and micro-level of issues in AKD, we propose a new KDD methodology on top of the traditional data-centered pattern mining

framework , that is *Domain Driven Data Mining* ($D^3M$) [2,4,5]. In the next section, we introduce the main idea of $D^3M$.

## 1.2 What Is Domain Driven Data Mining

### *1.2.1 Basic Ideas*

The motivation of $D^3M$ is to view KDD as AKD-based problem-solving systems through developing effective methodologies, methods and tools. The aim of $D^3M$ is to make AKD system deliver business-friendly and decision-making rules and actions that are of solid technical significance as well. To this end, $D^3M$ caters for the effective involvement of the following ubiquitous intelligence surrounding AKD-based problem-solving.

- *Data Intelligence* , tells stories hidden in the data about a business problem.
- *Domain Intelligence* , refers to domain resources that not only wrap a problem and its target data but also assist in the understanding and problem-solving of the problem. Domain intelligence consists of qualitative and quantitative intelligence. Both types of intelligence are instantiated in terms of aspects such as domain knowledge, background information, constraints, organization factors and business process, as well as environment intelligence, business expectation and interestingness.
- *Network Intelligence* , refers to both web intelligence and broad-based network intelligence such as distributed information and resources, linkages, searching, and structured information from textual data.
- *Human Intelligence*, refers to (1) explicit or direct involvement of humans such as empirical knowledge, belief, intention and expectation, run-time supervision, evaluating, and expert group; (2) implicit or indirect involvement of human intelligence such as imaginary thinking, emotional intelligence, inspiration, brainstorm, and reasoning inputs.
- *Social Intelligence* , consists of interpersonal intelligence, emotional intelligence, social cognition, consensus construction, group decision, as well as organizational factors, business process, workflow, project management and delivery, social network intelligence, collective interaction, business rules, law, trust and so on.
- *Intelligence Metasynthesis* , the above ubiquitous intelligence has to be combined for the problem-solving. The methodology for combining such intelligence is called *metasynthesis* [10, 11], which provides a human-centered and human-machine-cooperated problem-solving process by involving, synthesizing and using ubiquitous intelligence surrounding AKD as need for problem-solving.

### 1.2.2 $D^3M$ for Actionable Knowledge Discovery

Real-world data mining is a complex problem-solving system. From the view of systems and microeconomy, the endogenous character of actionable knowledge discovery (AKD) determines that it is an optimization problem with certain objectives in a particular environment. We present a formal definition of AKD in this section. We first define several notions as follows.

Let *DB* be a database collected from business problems ($\Psi$), $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L\}$ be the set of items in the *DB*, where $\mathbf{x}_l$ ($l = 1, \ldots, L$) be an itemset, and the number of attributes ($v$) in *DB* be $S$. Suppose $E = \{e_1, e_2, \cdots, e_K\}$ denotes the environment set, where $e_k$ represents a particular environment setting for AKD. Further, let $M = \{m_1, m_2, \cdots, m_N\}$ be the data mining method set, where $m_n$ ($n = 1, \ldots, N$) is a method. For the method $m_n$, suppose its identified pattern set $P^{m_n} = \{p_1^{m_n}, p_2^{m_n}, \cdots, p_U^{m_n}\}$ includes all patterns discovered in *DB*, where $p_u^{m_n}$ ($u = 1, \ldots, U$) denotes a pattern discovered by the method $m_n$.

In the real world, data mining is a problem-solving process from business problems ($\Psi$, with problem status $\tau$) to problem-solving solutions ($\Phi$):

$$\Psi \rightarrow \Phi \tag{1.1}$$

From the modeling perspective, such a problem-solving process is a state transformation process from source data $DB(\Psi \rightarrow DB)$ to resulting pattern set $P(\Phi \rightarrow P)$.

$$\Psi \rightarrow \Phi :: DB(v_1, \ldots, v_S) \rightarrow P(f_1, \ldots, f_Q) \tag{1.2}$$

where $v_s$ ($s = 1, \ldots, S$) are attributes in the source data *DB*, while $f_q$ ($q = 1, \ldots, Q$) are features used for mining the pattern set *P*.

**Definition 1.1.** (Actionable Patterns)
Let $\widetilde{P} = \{\tilde{p_1}, \tilde{p_2}, \cdots, \tilde{p_Z}\}$ be an *Actionable Pattern Set* mined by method $m_n$ for the given problem $\Psi$ (its data set is *DB*), in which each pattern $\tilde{p_z}$ is *actionable* for the problem-solving if it satisfies the following conditions:

1.a. $t_i(\tilde{p_z}) \geq t_{i,0}$; indicating the pattern $\tilde{p_z}$ satisfying technical interestingness $t_i$ with threshold $t_{i,0}$;
1.b. $b_i(\tilde{p_z}) \geq b_{i,0}$; indicating the pattern $\tilde{p_z}$ satisfying business interestingness $b_i$ with threshold $b_{i,0}$;
1.c. $R : \tau_1 \xrightarrow{A, m_n(\tilde{p_z})} \tau_2$; the pattern can support business problem-solving ($R$) by taking action $A$, and correspondingly transform the problem status from initially nonoptimal state $\tau_1$ to greatly improved state $\tau_2$.

Therefore, the discovery of actionable knowledge (AKD) on data set *DB* is an iterative optimization process toward the actionable pattern set $\widetilde{P}$.

$$AKD : DB \xrightarrow{e, \tau, m_1} P_1 \xrightarrow{e, \tau, m_2} P_2 \cdots \xrightarrow{e, \tau, m_n} \widetilde{P} \tag{1.3}$$

**Definition 1.2.** (Actionable Knowledge Discovery)
The *Actionable Knowledge Discovery* (AKD) is the procedure to find the *Actionable Pattern Set* $\widetilde{P}$ through employing all valid methods $M$. Its mathematical description is as follows:

$$AKD^{m_i \in M} \longrightarrow O_{p \in P} Int(p), \tag{1.4}$$

where $P = P^{m_1} U P^{m_2}, \cdots, UP^{m_n}$, $Int(.)$ is the evaluation function, $O(.)$ is the optimization function to extract those $\tilde{p} \in \widetilde{P}$ where $Int(\tilde{p})$ can beat a given benchmark.

For a pattern $p$, $Int(p)$ can be further measured in terms of *technical interestingness* ($t_i(p)$) and *business interestingness* ($b_i(p)$) [3].

$$Int(p) = I(t_i(p), b_i(p)) \tag{1.5}$$

where $I(.)$ is the function for aggregating the contributions of all particular aspects of interestingness.

Further, $Int(p)$ can be described in terms of *objective* ($o$) and *subjective* ($s$) factors from both *technical* ($t$) and *business* ($b$) perspectives.

$$Int(p) = I(t_o(), t_s(), b_o(), b_s()) \tag{1.6}$$

where $t_o()$ is objective technical interestingness, $t_s()$ is subjective technical interestingness, $b_o()$ is objective business interestingness, and $b_s()$ is subjective business interestingness.

We say $p$ is truly *actionable* (i.e., $\widetilde{p}$) both to academia and business if it satisfies the following condition:

$$Int(p) = t_o(\mathbf{x}, \widetilde{p}) \wedge t_s(\mathbf{x}, \widetilde{p}) \wedge b_o(\mathbf{x}, \widetilde{p}) \wedge b_s(\mathbf{x}, \widetilde{p}) \tag{1.7}$$

where $I \rightarrow$ '$\wedge$' indicates the 'aggregation' of the interestingness.

In general, $t_o()$, $t_s()$, $b_o()$ and $b_s()$ of practical applications can be regarded as independent of each other. With their normalization (expressed by $\hat{}$), we can get the following:

$$\begin{aligned} Int(p) &\rightarrow \hat{I}(\hat{t_o}(), \hat{t_s}(), \hat{b_o}(), \hat{b_s}()) \\ &= \alpha \hat{t_o}() + \beta \hat{t_s}() + \gamma \hat{b_o}() + \delta \hat{b_s}() \end{aligned} \tag{1.8}$$

So, the AKD optimization problem can be expressed as follows:

$$\begin{aligned} AKD^{e, \tau, m \in M} &\longrightarrow O_{p \in P}(Int(p)) \\ &\rightarrow O(\alpha \hat{t_o}()) + O(\beta \hat{t_s}()) + \\ &\quad O(\gamma \hat{b_o}()) + O(\delta \hat{b_s}()) \end{aligned} \tag{1.9}$$

**Definition 1.3.** (Actionability of a Pattern)
The *actionability* of a pattern $p$ is measured by $act(p)$:

$$\begin{aligned} act(p) &= O_{p\in P}(Int(p)) \\ \rightarrow O(\alpha\hat{t_o}(p)) &+ O(\beta\hat{t_s}(p)) + \\ O(\gamma\hat{b_o}(p)) &+ O(\delta\hat{b_s}(p)) \\ \rightarrow t_o^{act} &+ t_s^{act} + b_o^{act} + b_s^{act} \\ \rightarrow t_i^{act} &+ b_i^{act} \end{aligned} \tag{1.10}$$

where $t_o^{act}$, $t_s^{act}$, $b_o^{act}$ and $b_s^{act}$ measure the respective actionable performance in terms of each interestingness element.

Due to the inconsistency often existing at different aspects, we often find the identified patterns only fitting in one of the following sub-sets:

$$\begin{aligned} Int(p) \rightarrow \{\{t_i^{act}, b_i^{act}\}, \{\neg t_i^{act}, b_i^{act}\}, \\ \{t_i^{act}, \neg b_i^{act}\}, \{\neg t_i^{act}, \neg b_i^{act}\}\} \end{aligned} \tag{1.11}$$

where '$\neg$' indicates the corresponding element is not satisfactory.

Ideally, we look for actionable patterns $p$ that can satisfy the following:

*IF*

$$\begin{aligned} \forall p \in \widetilde{P}, \exists \mathbf{x} : t_o(\mathbf{x}, p) \wedge t_s(\mathbf{x}, p) \wedge b_o(\mathbf{x}, p) \\ \wedge b_s(\mathbf{x}, p) \rightarrow act(p) \end{aligned} \tag{1.12}$$

*THEN:*

$$p \rightarrow \widetilde{p}. \tag{1.13}$$

However, in real-world mining, as we know, it is very challenging to find the most actionable patterns that are associated with both 'optimal' $t_i^{act}$ and $b_i^{act}$. Quite often a pattern with significant $t_i()$ is associated with unconfident $b_i()$. Contrarily, it is not rare that patterns with low $t_i()$ are associated with confident $b_i()$. Clearly, AKD targets patterns confirming the relationship $\{t_i^{act}, b_i^{act}\}$.

Therefore, it is necessary to deal with such possible conflict and uncertainty amongst respective interestingness elements. However, it is a kind of artwork and needs to involve domain knowledge and domain experts to tune thresholds and balance difference between $t_i()$ and $b_i()$. Another issue is to develop techniques to balance and combine all types of interestingness metrics to generate uniform, balanced and interpretable mechanisms for measuring knowledge deliverability and extracting and selecting resulting patterns. A reasonable way is to balance both sides toward an acceptable tradeoff. To this end, we need to develop interestingness aggregation methods, namely the $I - function$ (or '$\wedge$') to aggregate all elements of interestingness. In fact, each of the interestingness categories may be instantiated into more than one metric. There could be several methods of doing the aggregation, for instance, empirical methods such as business expert-based voting, or more quantitative methods such as multi-objective optimization methods.

## 1.3 Open Issues and Prospects

To effectively synthesize the above ubiquitous intelligence in AKD-based problem-solving systems, many research issues need to be studied or revisited.

- Typical research issues and techniques in *Data Intelligence* include mining in-depth data patterns, and mining structured knowledge in unstructured data.
- Typical research issues and techniques in *Domain Intelligence* consist of representation, modeling and involvement of domain knowledge, constraints, organizational factors, and business interestingness.
- Typical research issues and techniques in *Network Intelligence* include information retrieval, text mining, web mining, semantic web, ontological engineering techniques, and web knowledge management.
- Typical research issues and techniques in *Human Intelligence* include human-machine interaction, representation and involvement of empirical and implicit knowledge.
- Typical research issues and techniques in *Social Intelligence* include collective intelligence, social network analysis, and social cognition interaction.
- Typical issues in *intelligence metasynthesis* consist of building metasynthetic interaction (m-interaction) as working mechanism, and metasynthetic space (m-space) as an AKD-based problem-solving system [6].

Typical issues in actionable knowledge discovery through m-spaces consist of

- Mechanisms for acquiring and representing unstructured and ill-structured, uncertain knowledge such as empirical knowledge stored in domain experts' brains, such as unstructured knowledge representation and brain informatics;
- Mechanisms for acquiring and representing expert thinking such as imaginary thinking and creative thinking in group heuristic discussions;
- Mechanisms for acquiring and representing group/collective interaction behavior and impact emergence, such as behavior informatics and analytics;
- Mechanisms for modeling learning-of-learning, i.e., learning other participants' behavior which is the result of self-learning or ex-learning, such as learning evolution and intelligence emergence.

## 1.4 Conclusions

The mainstream data mining research features its dominating focus on the innovation of algorithms and tools yet caring little for their workable capability in the real world. Consequently, data mining applications face significant problem of the workability of deployed algorithms, tools and resulting deliverables. To fundamentally change such situations, and empower the workable capability and performance of advanced data mining in real-world production and economy, there is an urgent need to develop next-generation data mining methodologies and techniques

that target the paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge discovery. Its goal is to build KDD as an AKD-based problem-solving system.

Based on our experience in conducting large-scale data analysis for several domains, for instance, finance data mining and social security mining, we have proposed the *Domain Driven Data Mining* ($D^3M$ for short) methodology. $D^3M$ emphasizes the development of methodologies, techniques and tools for *actionable knowledge discovery*. It involves relevantly ubiquitous intelligence surrounding the business problem-solving, such as human intelligence, domain intelligence, network intelligence and organizational/social intelligence, and the meta-synthesis of such ubiquitous intelligence into a human-computer-cooperated closed problem-solving system.

Our current work includes an attempt on theoretical studies and working case studies on a set of typically open issues in $D^3M$. The results will come into a monograph named *Domain Driven Data Mining*, which will be published by Springer in 2009.

**Acknowledgements** This work is sponsored in part by Australian Research Council Grants (DP0773412, LP0775041, DP0667060).

## References

1. Ankerst, M.: Report on the SIGKDD-2002 Panel the Perfect Pata Mining Tool: Interactive or Automated? ACM SIGKDD Explorations Newsletter, 4(2):110-111, 2002.
2. Cao, L., Yu, P., Zhang, C., Zhao, Y., Williams, G.: DDDM2007: Domain Driven Data Mining, ACM SIGKDD Explorations Newsletter, 9(2): 84-86, 2007.
3. Cao, L., Zhang, C.: Knowledge Actionability: Satisfying Technical and Business Interestingness, International Journal of Business Intelligence and Data Mining, 2(4): 496-514, 2007.
4. Cao, L., Zhang, C.: The Evolution of KDD: Towards Domain-Driven Data Mining, International Journal of Pattern Recognition and Artificial Intelligence, 21(4): 677-692, 2007.
5. Cao, L.: Domain-Driven Actionable Knowledge Discovery, IEEE Intelligent Systems, 22(4): 78-89, 2007.
6. Cao, L., Dai, R., Zhou, M.: Metasynthesis, M-Space and M-Interaction for Open Complex Giant Systems, technical report, 2008.
7. Fayyad, U., Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases, AI Magazine, 37-54, 1996.
8. 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.
9. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
10. Qian, X.S., Yu, J.Y., Dai, R.W.: A New Scientific Field–Open Complex Giant Systems and the Methodology, Chinese Journal of Nature, 13(1) 3-10, 1990.
11. Qian, X.S. (Tsien H.S.): Revisiting issues on open complex giant systems, Pattern Recognition and Artificial Intelligence, 4(1): 5-8, 1991.