Domain Driven Data Mining (D^3M)

Longbing Cao
Data Sciences and Knowledge Discovery Lab
Faculty of Engineering and Information Technology,
University of Technology, Sydney, Australia
lbcao@it.uts.edu.au

Abstract

In deploying data mining into the real-world business, we have to cater for business scenarios, organizational factors, user preferences and business needs. However, the current data mining algorithms and tools often stop at the delivery of patterns satisfying expected technical interestingness. Business people are not informed about how and what to do to take over the technical deliverables. The gap between academia and business has seriously affected the widespread employment of advanced data mining techniques in greatly promoting enterprise operational quality and productivity. To narrow down the gap, cater for real-world factors relevant to data mining, and make data mining workable in supporting decision-making actions in the real world, we propose the methodology of Domain Driven Data Mining *(D^3M for short). D^3M aims to construct next-generation methodologies, techniques and tools for a possible paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge delivery. In this talk, we address the concept map of D^3M, theoretical underpinnings, several general and flexible frameworks, research issues, possible directions, application areas etc. related to D^3M. Real-world case studies in financial data mining and social security mining are demonstrated to show the effectiveness and applicability of D^3M in both research and development of real-world challenging problems.*

1. What Is D^3M

Domain Driven Data Mining (D^3M for short) [1, 5, 4, 3, 2, 8, 13, 14, 15] targets the development of next-generation data mining methodologies, frameworks, algorithms, evaluation systems, tools and decision support, which aim to promote the paradigm shift from *data-centered hidden pattern mining* to *domain-driven actionable knowledge discovery (AKD)*. To this end, D^3M needs to involve and integrate human intelligence, domain intelligence, data intelligence, network intelligence, organizational and social intelligence, and the meta-synthesis of the above ubiquitous intelligence. As a result of the D^3M research and development, the AKD system can deliver business-friendly and decision-making rules and actions that are of solid technical and business significance.

2. Why Do We Need D^3M

In data mining community, 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.

From both macro-level and micro-level, we can find reasons asking for new methodology and paradigm shift such as domain driven data mining. On the macro-level, issues related to methodological and fundamental aspects include

- 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;

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.98

- If AKD is the target, we then have to cater for real-world aspects such as business processes, organizational factors, and constraints.

3. The D^3M Framework

D^3M advocates a framework of 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:

$$AKD^{m_i \in M} \longrightarrow O_{p \in P} Int(p), \tag{1}$$

where $P = P^{m_1}UP^{m_2}, \cdots, UP^{m_n}$, $Int(.)$ 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.

Correspondingly, the *actionability* of a pattern p is measured by $act(p)$:

$$\begin{aligned} act(p) &= O_{p \in P}(Int(p)) \\ &\rightarrow O(\alpha \hat{t_o}(p)) + O(\beta \hat{t_s}(p)) + \\ &\quad O(\gamma \hat{b_o}(p)) + O(\delta \hat{b_s}(p)) \\ &\rightarrow t_o^{act} + t_s^{act} + b_o^{act} + b_s^{act} \\ &\rightarrow t_i^{act} + b_i^{act} \end{aligned} \tag{2}$$

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.

4. D^3M Theoretical Underpinnings

D^3M involves and requires the theoretical foundations in areas such as knowledge representation and management, business modeling, business process management, statistics, machine learning, human-machine interaction, organizational and social computing, data integration, ontological engineering, social network analysis, system simulation, artificial intelligence and intelligent systems, behavior informatics and analytics, cognitive sciences, human-centered computing, project management methodology, and so on.

5. D^3M Research Issues

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 [**?**].

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.

6. D^3M Applications

Given the nature of D^3M, it can bring about the effective and practical development of many challenging data mining applications in every area. Based on the collaborations with our business partners, we have the experience in developing and deploying D^3M in areas such as capital markets and social security area. In capital markets, we develop actionable trading agents [10], actionable trading strategies [12], and

exceptional market microstructure behavior patterns [9]. In social security area, we propose the concept of activity mining [6] and combined mining [16, 12].

7. D^3M References

In the following sections, we list our works on D^3M. These papers involve the proposal, main ideas and typical applications of D^3M.

References

[1] Cao, L., and Zhang, C.Domain-driven data mining: A practical methodology, International Journal of Data Warehousing and Mining (IJDWM), IGI Global, 2(4): 49-65, 2006.

[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., Zhao, Y., Zhang, C. (2008), Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. Knowledge and Data Engineering, IEEE, , Vol. 20, No. 8, pp. 1053-1066, 2008..

[7] Cao, L., Dai, R., Zhou, M.: Metasynthesis, M-Space and M-Interaction for Open Complex Giant Systems, technical report, 2008.

[8] Cao, L. Yu, P.S., Zhang, C., Zhang, H. Data Mining for Business Applications, Springer, 2008.

[9] Cao, L. and Ou, Y. Market Microstructure Patterns Powering Trading and Surveillance Agents. Journal of Universal Computer Sciences, 2008 (to appear).

[10] Cao, L. and He, T. Developing actionable trading agents, Knowledge and Information Systems: An International Journal.

[11] Cao,L. Developing Actionable Trading Strategies, in edited book: Intelligent Agents in the Evolution of WEB and Applications, Springer.

[12] Cao, L., Zhang, H., Zhao, Y., Zhang, C. Combined Mining for More Informative Knowledge in e-Government Services, technical report, 2008.

[13] Cao, L. Introduction to Domain Driven Data Mining, Data Mining for Business Applications (Edited Book), Springer, 2008.

[14] Cao, L. Actionable Knowledge Discovery, Encyclopedia of Information Science and Technology, Second Ed., IGI Global publications, 2008.

[15] Cao, L. Yu, P.S., Zhang, C., Zhao, Y. Domain Driven Data Mining, Springer, 2009.

[16] Zhao, Y., Zhang, H., Cao, L., Zhang, C. and Bohlscheid, H. Combined pattern mining: from learned rules to actionable knowledge. Proc. of the Twenty-First Australasian Joint Conference on Artificial Intelligence (AI 08), 2008.