Why Data Mining Research Does Not Contribute to Business?

Mykola Pechenizkiy[1], Seppo Puuronen[1], Alexey Tsymbal[2]

[1]Dept. of Computer Science and Inf. Systems, University of Jyväskylä, Finland
{mpechen,sepi}@cs.jyu.fi
[2]Dept. of Computer Science, Trinity College Dublin, Ireland
alexey.tsymbal@cs.tcd.ie

Abstract. Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it. Nowadays there exist many DM algorithms, developed, implemented and available for direct use or integration into specific solution. There exist also a number of DM systems that provide DM tools for all steps of the DM process. This paper is aimed at provoking the discussion - why contribution of data mining research to business has not been as powerful as was expected. We strongly emphasize that practice-oriented aspects require more attention in DM research, and we stress that DM research should take into account the practice relevance of research beside the scientific rigor of it. This might require broadening the current narrow DM research framework.

1 Introduction

Business organizations due to recent advances in information and database technologies are able to collect vast amount of data that potentially can be processed and used in decision-making processes. Data mining (DM) and knowledge discovery are intelligent tools that help to process data and make use of it [4].

Nowadays there exist a number of DM algorithms to identify "valid, novel, potentially useful, and ultimately understandable patterns in data" [4]. In a data mining systems (DMS), as a rule, fairly many of those techniques are implemented and are available for accomplishing DM tasks. However, despite the maturity of the field in this sense recent discussions state that in fact DM (DM systems and DM solutions) does not contribute to business in a large scale as for example DBMS systems do.

What might be a reason for that? SPSS whitepaper [2] states that "Unless there's a method, there's madness". It is accepted that just by pushing a button someone should not expect useful results to appear. An industry standard to DM projects CRISP-DM is a good initiative and a starting point directed towards the development of DM meta-artifact (methodology to produce DM artifacts). However, in our opinion it is just one guideline, which is in very general-level offering a DM developer weak guidelines to follow

successfully. Process-oriented frameworks try to address the iterativeness and interactiveness of the DM process. However, the development process of a DM artifact and the use of it is poorly (if ever) emphasized.

Someone may claim as e.g. Lin in Wu et al. [7] that the research and development goals of DM are quite different, since research is a knowledge-oriented effort while development is a profit-oriented one. Thus, according them DM researchers are mainly concentrated on the development of new algorithms or their enhancements without good connections to application domains and their restrictions. On the other hand the DM developers in domain areas are aware of cost and benefit considerations as investment in research, product development, marketing, and product support. We agree that this clearly describes the current state of the DM field. However, we believe that the study of the DM development and DM use processes is becoming equally important now when many technological aspects within DM research have been solved at least in usable level. This does not exclude the importance of further technological DM research.

2 Rigor and Relevance in Data Mining Research

DM researchers, having Statistics, Machine Learning, Artificial Intelligence, Databases and other disciplines as reference disciplines contributing to DM, naturally emphasize the "rigor" aspects of the research. This is likely the same main reason for the fact that currently, DM research rarely, if ever, considers many important "relevance" issues, including the external environment (legal, social, political, cultural, economic, educational, resource and industry/trade considerations), the organizational environment marked by the organizational goals, tasks, structure, and management, and the operations environment that incorporates the resources necessary for DM operations (software, hardware, database, procedures/ documentation, organization and management of DM operations).

Many research paradigms have been suggested and used with this respect in the Information Systems (IS) discipline. Currently, Hevner et al. [6] suggest that two paradigms should be recognized within research in the IS discipline. These are the behavioural-science paradigm and the design-science paradigm. According to the Hevner et al., the behavioural science paradigm tries "to develop and verify theories that explain or predict human or organizational behaviour". This paradigm is naturally the most broadly applied in the IS use process related topics. They continue that "The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts". This second paradigm is the most natural in the IS development related topics where the new user and development environments are planned and experimented with. Some others call the IS development process related research as a constructive type of research since development always involves creation of some new artifacts – conceptual (models, frameworks) or more technical ones (as software implementations).

3 DM Success Model

Investments into IT, information systems are huge every year and have been so already decades. It has resulted to the natural needs to evaluate the effectiveness of IS expenditures and thus research interest in this area. A similar situation can be seen developing with DM when DM techniques are crawling from the research laboratories to business organizations. It is surprising that so little research have been accomplished with the DM success model construction even when these kind of models typically raise information quality, service quality, and systems quality as key ingredients behind the user satisfaction and the use of systems. At least, these have been found to have essential positive effect to individual impact leading to the organizational impact in information systems research (IS) [3]. A similar approach to that with IS is needed with DMS to recognize the key factors of successful use and impact of DMS both at the individual and organizational levels. Questions like (1) how the system is used, and also supported and evolved, and (2) how the system impacts and is impacted by the contexts in which it is embedded are important also in the DMS context. The first efforts in that direction are the ones presented in the DM Review magazine [9, 21], referred below. We believe that such efforts should be encouraged in DM research and followed by research-based reports.

Coppock [1] analyzed, in a way, the failure factors of DM-related projects. In his opinion they have nothing to do with the skill of the modeler or the quality of data. But those do include these four: (1) persons in charge of the project did not *formulate actionable insights*, (2) the sponsors of the work did not *communicate the insights* derived to key constituents, (3) the results *don't agree with institutional truths*, and (4) the project never had a *sponsor and champion*. The main conclusion of Coppock's analysis is that, similar to an IS, the leadership, communication skills and understanding of the culture of the organization are not less important than the traditionally emphasized technological job of turning data into insights.

Hermiz [5] communicated his beliefs that there are four critical success factors for DM projects: (1) having a clearly articulated business problem that needs to be solved and for which DM is a proper tool; (2) insuring that the problem being pursued is supported by the right type of data of sufficient quality and in sufficient quantity for DM; (3) recognizing that DM is a process with many components and dependencies – the entire project cannot be "managed" in the traditional sense of the business word; (4) planning to learn from the DM process regardless of the outcome, and understand-ing, that there is no guarantee that any given DM project will be successful.

4 Do We Need a New Framework for DM Research?

Piatetsky-Shapiro in Wu *et al.* [7] gives a good example that characterizes the whole area of current DM research: "we see many papers proposing incremental refinements in association rules algorithms, but very few papers describing how the discovered

association rules are used". DM is a fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data [7]. A DMS is generally used to support or do some task(s) of people in an organizational environment. Beside organizational goals those individual actors have their own desires related to the DMS. Further, the organization has its own environment that has its own interest related to the DMS, for example that privacy of people is not violated.

Thus, the environment defines not only the data that represents the problem to be mined but at least as importantly people, business organizations, and their existing or desired technologies, infrastructures and development capabilities. Those include business goals, tasks, problems, and opportunities that define business needs, which are assessed and evaluated within the context of organizational strategies, structure, culture, and existing business processes. The DM research activities that are aimed at addressing those business needs contribute to the relevance of DM research.

These key ideas are depicted in the very compressed form in Figure 1, where we try to adapt a conceptual framework for understanding, conducting and evaluation of the IS research (presented Heavner et al. [6]) to the context of DM research. The framework combines together the behavioral-science and design-science paradigms and shows how research rigor and research relevance can be explained, evaluated, and balanced.

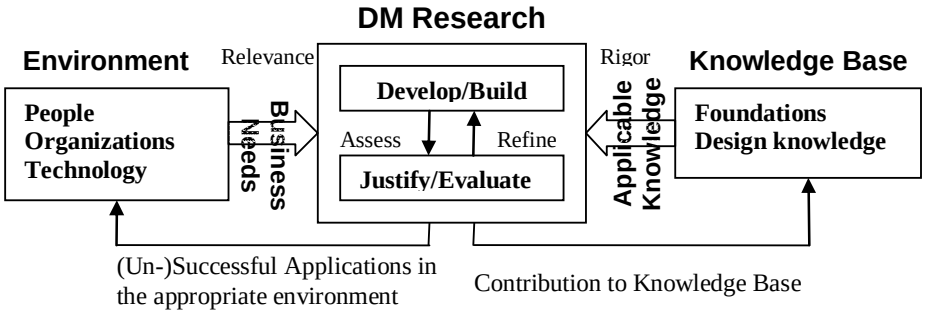

Fig. 1. A new research framework for DM research

Driven by the business needs, DM research can be conducted in two complementary phases as it is proposed in [6] for the IS research. Behavioral science would guide research through the *development* and *justification* of theories that describe, explain or predict some phenomena associated with the business need(s) being addressed. Design science enables the *building* and *evaluation* of artifacts being developed to address the business need(s). Prior DM research and development, and results from reference disciplines (statistics, machine learning, AI, etc.) provide foundational theories, frameworks, models, methods, techniques and their instantiations used in the develop/build phase of research. Rigor is achieved by appropriately applying existing foundations and methodologies.

It is excellent that there exist rigor research results and they will be needed more and more in future. On the other hand we see DM research mature enough to be seriously

considered also by those who are interested in research that is able to take into account the relevance aspects of applying and using DM technologies in real life domains.

5 Conclusions

It was noticed by some researchers that in fact there have been no major impacts of DM on the business world echoed. In this paper we presented our vision of why it is so and considered the need for new DM research framework, which would aim at better balancing between the rigor and relevance of research.

Although we neither provided any examples nor presented any concrete guidelines how to improve the situation, we believe that our work could be helpful in turning the focus of DM research into a better balanced direction. We see this important from the point of view of raising DM first among those technologies which are able to produce competitive advantage and later to be developed to be the one of everyday mainline technologies.

We hope that our work could raise a new wave of interest to the analysis of the DM field from different perspectives, including such important topics as DM success, DM costs, DM risks, DM life cycles, methods for analyzing systems, organizing and codifying knowledge about DM systems in organizations, and maximizing the value of DM research.

Acknowledgments: This material is based upon works partly supported by the COMAS Graduate School of the University of Jyväskylä, Finland and the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

References

1. Coppock D. S. “Data Mining and Modeling: So You have a Model, Now What?” *DM Review Magazine*, Feb 03.
2. CRISP-DM: 1.0 *Step-by-step DM guide*, SPSS Inc.
3. DeLone W., McLean E.R. “The DeLone and McLean Model of Information Systems Success: A Ten-Year Update”, *Journal of MIS 19(4)*, 2003, pp. 9-30
4. Fayyad U.M. “Data Mining and Knowledge Discovery: Making Sense Out of Data”, *IEEE Expert 11(5)*, 1996, pp. 20-25
5. Hermiz K.B., “Critical Success Factors for Data Mining Projects”, *DM Review Magazine*, February 1999.
6. Hevner A., March S., Park J., Ram S. Decision Science in Information Systems Research, MIS Quarterly 26(1), 2004, pp. 75-105.
7. Wu X., Yu P., Piatetsky-Shapiro G., et al. “Data Mining: How Research Meets Practical Development?” *Knowledge and Information Systems 5(2)*, 2000, pp. 248 – 261.