Competitive advantage from Data Mining: some lessons learnt in the Information Systems field

Mykola Pechenizkiy
Dept. of CS and ISs
University of Jyväskylä
Finland
mpechen@cs.jyu.fi

Seppo Puuronen
Dept. of CS and ISs
University of Jyväskylä
Finland
sepi@cs.jyu.fi

Alexey Tsymbal
Dept. of CS
Trinity College Dublin
Ireland
tsymbalo@tcd.ie

Abstract

Data mining (DM) is still a technology having great expectations to enable organizations to take more benefit of their huge databases. There exist some success stories where organizations have managed to have competitive advantage of DM. Still the strong focus of most DM-researchers in technology-oriented topics does not support expanding the scope in less rigorous but practically very relevant sub-areas. The current situation with DM has similarities with situations during the development of some other information technology (IT)-related sub-areas earlier. Research in the Information Systems discipline (one of those IT-related sub-areas) has strong traditions to take into account human and organizational aspects of systems beside the technical ones. We suggest in this paper that these user and organization related research results and organizational settings include essential points of view with respect to developing DM research to produce practically more relevant results for domain areas where human and organizational things matter.

1. Introduction

Data mining (DM) and knowledge discovery are commonly seen as intelligent tools that help to accumulate and process data and make use of it [9]. DM bridges many technical areas, including databases, statistics, machine learning, and human-computer interaction. The set of DM processes used to extract and verify patterns in data is the core of the knowledge discovery process [9].

Technical aspects of DM have received good amount of rigor research efforts and are maturing fast as one of the most potential new approaches to exploit large databases. Some companies have had and many more are planning to have pilot DM projects. An excellent collection of DM-algorithms and bright data miners are needed to implement those DM projects. But this is not enough for organizations to take full competitive advantage from DM. The problems considered and the solutions developed need to be selected carefully to support other efforts of the organization. Currently the maturation of DM-supporting processes which would take into account human and organizational aspects is still living its childhood.

In this paper we underline that the young DM community might benefit, at least from the practical point of view, looking at some other older sub-areas of IT having traditions to consider solution-driven concepts with a focus also on human and organizational aspects. One such challenging discipline to consider is Information Systems (IS). The DM community by becoming more amenable to research results of the IS community might be able to increase its collective understanding of (1) how DM artifacts are developed – conceived, constructed, and implemented, (2) how DM artifacts are used, and also supported and evolved, and (3) how DM artifacts impact and are impacted by the contexts in which they are embedded.

Nevertheless, so far in the DM community there exist too few research activities directed towards the study of a *DM system* as an *artifact* aimed to enable certain DM tasks in a certain context (Figure 1). In the IS discipline two research paradigms – the behavioral-science paradigm and design-science paradigm – have

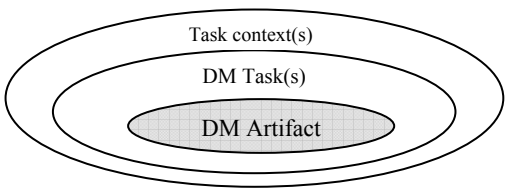

Figure 1. DM artifact (adapted from [1])

strong traditions in non-technology-oriented research topics, including human and organizational aspects of IS.

The rest of the paper is organized as follows. In Section 2 we (1) review the existing types of DM frameworks that emphasize different perspectives of DM, and (2) analyze the current state of DM development and suggest the possible direction of further development. In Section 3 we present a brief overview of the areas of IS and IS development. In Section 4 we address the issues of (1) DM artifact development and (2) use of DM artifact, trying to raise potential connections between the IS discipline and on-going initiatives in the DM community. We conclude briefly with a summary and some recommendations about how to make the use of findings from the IS field possible.

2. DM frameworks and development

The idea of learning from data is far from being new. However, likely due to developments in the database management and the huge increase of data volumes being accumulated in databases the interest in DM has become very intense. Numerous DM algorithms have recently been developed to extract knowledge from large databases. Currently, most research in the DM area focuses on the development of new algorithms or the improvement of speed or accuracy of the existing ones [21].

There is, although relatively little, on-going research directed towards the development of DM frameworks. Perhaps the first summary of a few theoretical approaches in the DM area was proposed in [17]. In the next subsection we present our brief overview of different DM frameworks.

2.1. Existing DM Frameworks

2.1.1. Theory-oriented frameworks. Frameworks of this type are based mainly on one the following paradigms: (1) the *statistical paradigms*: "Statistical experiment" paradigm (Fisher's version that uses the inductive principle of maximum likelihood, Neyman-E.S. Pearson-Wald's version that is based on the principle of inductive behavior, and the Bayesian version that is based on the principle of maximum posterior probability); an evolved version of the "Statistical experiment" paradigm that is "Statistical learning from empirical process" paradigm; and "Structural data analysis" that is associated with singular value decomposition methods; (2) *the data compression paradigm* – "compress the dataset by finding some structure or knowledge for it", where knowledge is interpreted as a representation that allows coding the data using less bits (e.g., the MDL principle [18] can be used to select among different encodings accounting to both the complexity of a model and its predictive accuracy); (3) *the machine learning paradigm* – "let the data suggest a model" that can be seen as a practical alternative to the statistical paradigm "fit a model to the data"; (4) *the database paradigm* – "there is no such thing as discovery, it is all in the power of the query language" [14]; and also *the inductive databases paradigm* – "locating interesting sentences from a given logic that are true in the database" [2].

2.1.2 Process-oriented frameworks. Frameworks of this type are known mainly because of works [9] and [4]. They view DM as a sequence of iterative processes that include data cleaning, feature transformation, algorithm and parameter selection, and evaluation, interpretation and validation.

SPSS whitepaper [4] 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 too general-level, that every DM developer follows with or without success.

2.1.3. Foundations-oriented frameworks. Some DM researchers argue for the lack of accepted fundamental conceptual framework or a paradigm for DM research and consequently for the need of some consensus on the fundamental concepts. Therefore, they try to search for some mathematical bricks for DM. One interesting approach based on granular and rough computing can be found in [16]. However, others may think that the current diversity in theoretical foundations and research methods is a good thing and also it might be more reasonable to search for an umbrella-framework that would cover the existing variety.

Another direction of research could lie in addressing data to be mined, DM models, and reality views through the prism of the philosophy of science paradigm, that includes consideration of nominalistic *vs* realistic ontological beliefs, voluntaristic *vs* deterministic assumptions about the nature of every instance constituting the observed data, subjectivist *vs* objectivist approaches to model construction, ideographic *vs* nomothetic view at reality; and epistemological assumptions about how a criterion to validate knowledge discovered can be constructed.

2.2. Where DM is and where to go

Different frameworks account for different DM tasks like clustering, regression, and classification. One way or another, we can easily see the exploratory nature of the frameworks for DM. It is agreed also that most approaches are lacking the ways for taking the iterative and interactive nature of the DM process into account [17]. In [20] we considered IS development and knowledge management perspectives emphasizing DM as a set of iterative and interactive processes.

Recently the focus has been on speeding-up, scaling-up, and increasing the accuracies of techniques/algorithms/methods. The microeconomic view on DM [15] is one good exception from this trend.

Although Dunkel *et al.* [8] concluded that there is a need and opportunity for computing systems research and development, almost 9 years later, to the best of our knowledge there are no significant research papers published in this direction in the DM area.

Lin in Wu *et al.* [21] notices that a new successful industry (as DM) can follow consecutive phases: (1) discovering a new idea, (2) ensuring its applicability, (3) producing small-scale systems to test the market, (4) better understanding of new technology and (5) producing a fully scaled system. At the present moment there are several dozens of DM systems, none of which can be compared to the scale of a DBMS system. This fact according to Lin indicates that we are still in the 3rd phase in the DM area.

Further in [21], Lin claims that the research and development goals of DM are quite different, since research is knowledge-oriented while development is profit-oriented. Thus, DM research is concentrated on the development of new algorithms or their enhancements but the DM developers in domain areas are aware of cost considerations: 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 equally important as the technological aspects and therefore such research activities are likely to emerge *within* the DM field. In fact, the study of development and use processes was recognized to be of importance in the IS fields many years ago, and therefore it has been introduced into the different IS frameworks.

3. An IS research framework and paradigms

ISs are powerful instruments for organizational problem solving through formal information processing. The traditional IS research framework presented in Figure 2 [5] is widely known in the IS community. In this framework an IS is considered in its organizational environment that is further surrounded by an external environment. According to this framework an IS itself includes three environments: a user environment, an IS development environment, and an IS operations environment. There are accordingly three processes through which an IS has interaction with its environments: the use process, the development process, and the operation process.

The research framework is thus very broad resulting in various different research questions and settings. The most extensive ones relate to the effects of IS into its organizational and external environments.

Hevner *et al.* [12] recognize two paradigms within the research in the IS discipline. These are the behavioural-science paradigm and the design-science paradigm. According to the authors, the behavioural science paradigm tries "to develop and verify theories that explain or predict human or organizational behaviour" [12, p. 75]. This paradigm is naturally the most broadly applied in the use process related topics. They continue [12, p. 75] that "The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts".

Some others as e.g. Iivari *et al.* [13] relate the IS development process to the constructive type of research based on the philosophical belief that development always involves creation of some new artifacts – conceptual (models, frameworks) or more technical artifacts (software implementations). They classify research as constructive whereas scientific knowledge is used to produce either useful systems or methods, including development of prototypes and processes. It has been argued that the constructive type of research is important especially for applied disciplines of IS and computer science [13], and DM may also be considered as such a discipline.

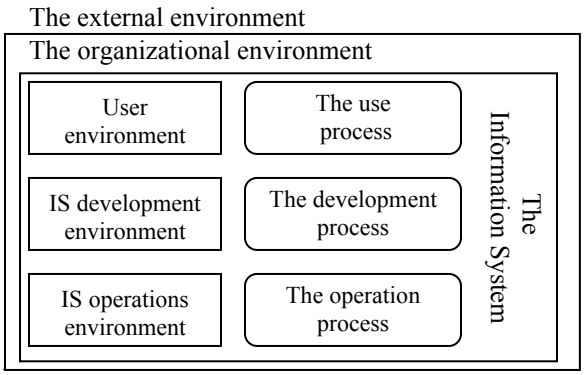

Figure 2. A framework for IS research [5]

Analogously, a data-mining system (DMS) can be considered as a system having organizational and external environments and including a user environment, a DM development environment, and a DM operations environment (as an IS in Figure 2).

4. The IS field as a new potential reference discipline for DM

In this section we consider a DMS as an IS including a number of techniques to be applied for a problem at hand using existing database(s). We consider the DMS development process and the DMS system use process leaving the less important operations environment process out of discussions.

4.1. The DMS development process

Nunamaker *et al.* [19] consider system development as a central part of a multi-methodological IS research cycle (Figure 3). Theory building involves the discovery of new knowledge in the field of study, however it rarely contributes directly to practice. Nevertheless, the new theory often (if not always) needs to be experimented with in the real world to check its validity, recognize its limitations and make refinements according to observations made during its application. Therefore the research methods can be subdivided into basic and applied research, as naturally both are common for any large system development project [19]. The proposed theory leads to the development of a prototype system in order to illustrate the theoretical framework on the one hand, and to test it through experimentation and observation with subsequent refinement of the theory and the prototype in an iterative manner. This view presents a research framework as a complete, comprehensive and dynamic DMS development process. It allows multiple perspectives and flexible choices of methods to be applied at different stages of the research process.

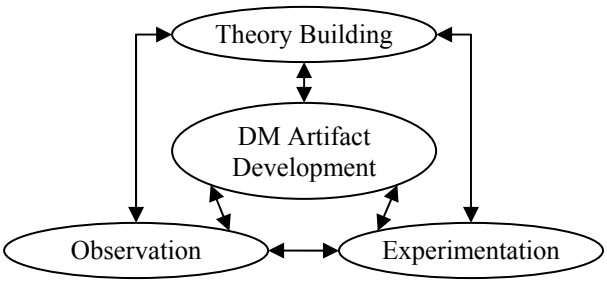

Figure 3. A multimethodological approach to the construction of an artifact for DM (adapted from [19])

4.2 The DMS use process

Piatetsky-Shapiro in Wu *et al.* [21, p. ..] gives a good example that characterizes the whole area of 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 fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data [21]. It is essential to make research related to the use processes of DMS, considering impacts and the essential factors that effect the impacts. One well-known success model in the IS discipline is presented in Figure 4.

A similar approach is needed with DMS to recognize the key factors of successful use and impact of DMS both in individual and organizational levels. The first efforts to that direction are the ones presented in the DM Review magazine [3,11] referred below and those should be followed by research-based reports.

Coppock [3] analyzed, in a way, the failure factors of DM-involved 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 as in an IS the leadership, communications skills and an understanding of the culture of the organization are not less important than the traditionally emphasized technological job of turning data into insights.

Hermiz [11] communicated his beliefs that there are the 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;

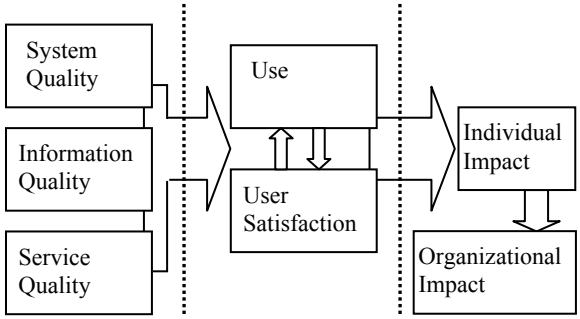

Figure 4. Adapted from D&M IS Success Model [7, p.87] and updated D&M IS Success Model [6, p.24]

(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 clearly understanding, that there is no guarantee that any given DM project will be successful.

Lin in Wu *et al.* [21] notices that in fact there have been no major impacts of DM on the business world echoed. However, even reporting of existing success stories is important. Giraud-Carrier [10] reported 136 success stories of DM, covering 9 business areas with 30 DM tools or DM vendors referred. Unfortunately, there was no deep analysis provided that would summarize or discover the main success factors and the research should be continued.

5. Conclusions

In this paper, first, we have considered briefly the history of development of DM. Then, after a brief overview of the IS area we considered more deeply the use and development- oriented sub-areas of IS. We suggested importing research problems and settings from the IS discipline to the DM research, for example towards having more DM research that uses IS as a reference discipline. 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 developed to be the one of the everyday mainline technologies.

We hope that our work could raise a new wave of interest to the foundations of DM and to the analysis of the DM field from different perspectives, similar to IS and ISD. This can be achieved by the building of knowledge networks across the field boundaries (DM and IS), e.g. by organizing a workshop that would include 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.

Acknowledgements. This research is partly supported by the COMAS Graduate School of the University of Jyväskylä, the Academy of Finland, and by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

6. References

[1] Benbasat I., Zmud R. W. "The Identity Crisis Within The IS Discipline: Defining and Communicating the Discipline's Core Properties", *MIS Quarterly 27(2)*, 2003, pp. 183-194.

[2] Boulicaut J., Klemettinen M., and Mannila H. "Modeling KDD Processes within the Inductive Database Framework". In *Proc. of the 1st DaWaK*, Springer, 1999, pp. 293-302.

[3] Coppock D. S. "Data Mining and Modeling: So You have a Model, Now What?" *DM Review Magazine*, Feb 03.

[4] CRISP-DM: 1.0 *Step-by-step DM guide*, SPSS Inc.

[5] Davis, G. "Information systems conceptual foundations: looking backward and forward", *Organizational and Social Perspectives on Information Technology*, R. Baskerville, J. Stage, and J. DeGross, (eds.), Kluwer, Boston, 2002.

[6] 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

[7] DeLone W., McLean E.R. "Information Systems Success: The Quest for the Dependent Variable", *Information Systems Research 3(1)*, 1992, pp. 60-95.

[8] Dunkel B.; Soparkar N.1; Szaro J.; Uthurusamy R. "Systems for KDD: From concepts to practice", *Future Generation Computer Systems 13(2)*, 1997, pp. 231-242

[9] Fayyad U. "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert 11(5)*, 1996, pp.20-25

[10] Giraud-Carrier C. *Success Stories in Data/Text Mining*, Brigham Young University, 2004 (An updated version of an ELCA Informatique SA White Paper)

[11] Hermiz K.B., "Critical Success Factors for Data Mining Projects", *DM Review Magazine*, February 1999.

[12] Hevner A., March S., Park J., Ram S. Decision Science in Information Systems Research, MIS Quarterly 26(1), 2004, pp. 75-105.

[13] Iivari J., Hirscheim R., Klein H. "A paradigmatic analysis contrasting information systems development approaches and methodologies", *Information Systems Research 9(2*), 1999, pp. 164-193

[14] Imielinski T., Mannila H. "A database perspective on knowledge discovery", *Communications of the ACM 39(11)*, 1996, pp. 58-64.

[15] Kleinberg J., Papadimitriou C., and Raghavan P. "A Microeconomic View of Data Mining," *Data Mining and Knowledge Discovery 2(4)*, 1998, pp. 311-324.

[16] Lin T.Y. "Data Mining: Granular Computing Approach" *Proc. 3rd PAKDD*, LNCS 1547, 1999, pp. 24-33.

[17] Mannila H. "Theoretical Framework for Data Mining", *SIGKDD Explorations 1(2)*, 2000, pp. 30-32.

[18] Mehta M., Rissanen J., Agrawal R. "MDL-Based Decision Tree Pruning", In *Proc. KDD'95*,1995, pp. 216-221

[19] Nunamaker W., Chen M., Purdin T. "Systems development in information systems research", *Journal of Management Information Systems 7(3)*, 1990-91, 89-106.

[20] Pechenizkiy M., Puuronen S., Tsymbal A. "The iterative and interactive data mining process: the ISD and KM perspectives" *Proc. Foundations of Data Mining Workshop*, 2004, pp. 129-136.

[21] Wu X., Yu P., Piatetsky-Shapiro G., et al. "Data Mining: How Research Meets Practical Development?" *Knowledge and Inf. Systems 5(2)*, 2000, pp. 248 – 261.