Data Mining Researcher, Who is Your Customer? Some Issues Inspired by the Information Systems Field

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

Mykola Pechenizkiy
Dept. of Mathematical IT
University of Jyväskylä
Finland
mpechen@cs.jyu.fi

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

Abstract

Data mining as an applied research field is still causing great expectations among organizations which want to raise the utility they are getting from their huge databases and data warehouses. There exist too few success stories about organizations having managed to satisfy even some of those expectations. This situation is very similar to the one inside the information systems (IS) field, especially earlier but even currently. The recent lively debate about the identity of the IS discipline included also the analysis concerning the customers of IS research. Inspired by IS researchers' insights related to the topic, we ask the question "who is our customer?" as data mining researchers. With this we want to raise to discussion the border that limits the topics 'acceptable' to work with as a data mining researcher. We suggest in this paper that the border should be transferred more clearly towards the direction so that beside the technical concerns also at least some user- and organization-related research questions are included.

1. Introduction

Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it [11]. They bridge many technical areas, such as databases, statistics, machine learning, and human-computer interaction. The set of DM processes used to extract and verify patterns in data is the hard core of the knowledge discovery process [11].

Technical aspects of DM have received good amount of rigor research efforts and are maturing fast, demonstrating a huge potential in exploiting existing large data bases. 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 these 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, too. Currently the maturation of DM-supporting processes which would take into account human and organizational aspects is still living its childhood.

There has been quite much research dedicated to DM frameworks. In [15] we presented a comprehensive review of existing DM frameworks grouping them into theory-oriented, process-oriented, and foundation-oriented categories.

The theory-oriented frameworks are based mainly on one of the following paradigms: (1) *the three statistical paradigms*: statistical experiment paradigm, statistical learning from empirical process paradigm, or structural data analysis paradigm, (2) *the data compression paradigm* where the dataset is compressed by finding some structure or knowledge for it, (3) *the machine learning paradigm* where the idea is to let the data suggest a model, and (4) *the database paradigm* based on the idea that all the power of discovery is in the query language.

The process-oriented frameworks view DM as a sequence of interactive processes that include data cleaning, feature transformation, algorithm and parameter selection, and evaluation, interpretation, and validation. CRISP-DM [6] is maybe the best example of the methodology for DM artifact production.

The foundation-oriented frameworks are based on the idea that DM research needs a commonly accepted conceptual framework or a paradigm in order to form consensus on fundamental concepts. There are also strong opinions supporting diversity seeing an umbrella-framework as a more reasonable one. A similar kind of discussion about the core of IS research has been going on quite a while, see for example two recent works [1,2].

Different frameworks account for different DM tasks like clustering or classification. These raise the exploratory nature of the frameworks for DM, but still there are too few approaches taking utility into account [15].

In this paper we focus on considering *the stakeholders of DM research* and on discussing which ones of them

Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)
0-7695-2641-1/06 $20.00 © 2006 IEEE

should be considered as the customers of DM research. The question '*who is a customer of research?*' has earlier been discussed in other research areas, such as the Information System (IS) discipline that we refer to. (An IS can be seen here as an instrument for organizational problem solving through formal information processing). It is important, we think, that in the IS community an IS is often considered (in the traditional IS research framework [8]) in its organizational environment that is surrounded by an external environment.

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).

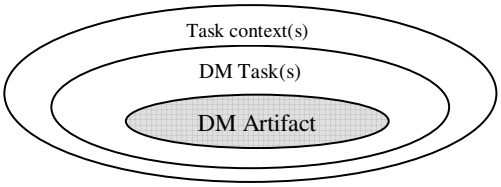

Figure 1. DM artifact (adapted from [2])

The rest of the paper is organized as follows. In Section 2 we discuss about the stakeholders of research, especially DM research. In section 3 we continue our discussion considering the relevance aspect of research from the different stakeholders' point of view. In Section 4 we address the issues of DM artifact development and use, and some points of view broadening the vision beyond artifacts. We conclude briefly with some remarks.

2. Stakeholders of DM research

Under *a stakeholder* of DM research we mean a person or organization that has a legitimate interest in DM research or its results. We divide the stakeholders of DM research into two groups: 1) *internal stakeholders* that are stakeholders within academia, and 2) *external stakeholders* that are all the others outside academia. Under *a customer* of DM research we mean a stakeholder that makes use of the results of DM research.

Related to IS research, Hirschheim and Klein [13, p.249] stress the need to recognize its stakeholders raising a question "who the stakeholders for our research are and what relevancy means for them". They mention as external stakeholders of publicly-funded research the following: industry shareholders and their agents (management), the employees of firms and organizations, their agents (unions), community and other levels of government and the general public. Beside external stakeholders they refer to [3] that IS researchers have important stakeholders within academia, as funding agencies, colleagues in other disciplines, university administrators, and students.

In their detailed analyzes of external stakeholders they [13, p.250] focus on the most commonly espoused group, the industry management considering two subgroups: senior management and the practitioners in IS departments. From the internal stakeholders they consider also two groups: IS research community and academics in other disciplines.

Hevner *et al*. [12] have considered the design science setting from the relevance and rigor point of view. Figure 2 is simplified and modified version of their original figure applied to DM research. In this figure we have on the left-hand side the environment of the DM research which includes stakeholders having business needs related to DM research. On the right-hand side in the figure is the knowledge base related to the DM research. It forms the base for rigorous research and it can be thought as an internal stakeholder expecting contributions from DM research. Hevner *et al*. [12] support the idea that IS research at the best serve both the environment and the knowledge base.

Chiasson and Davidson [4] considered various ways industry can be addressed in IS research and they assessed how industry influences IS activities. They found that industries often provide important contexts and so-called *contextual spaces* to build a new theory and to refine or evaluate the boundaries of existing theory.

There are opposite opinions also in the IS research area, as Alter [1] mentions referring to the rigor vs. relevance discussions [7] of IS research. According to Alter [1, p.503], "the IS academic community is the customer of academically respectable IS research; publications written to be understandable and usable by practitioners are often viewed as unworthy of credit within the academic community". He further raises a broader question about the customers of the IS discipline, not just IS research publications.

Lin in Wu *et al.* [18] 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 Lin's claim clearly describes the current state of most DM research. However, we want to raise the question, "Is it reasonable that DM researchers leave the study of the DM development and DM use processes totally outside their research area?". Are these equally important aspects going to be handled better by the researchers of other areas, and DM researchers should also in the future concentrate on the technological aspects of DM only?

In any case it is evident that DM research has both

Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)
0-7695-2641-1/06 $20.00 © 2006 IEEE

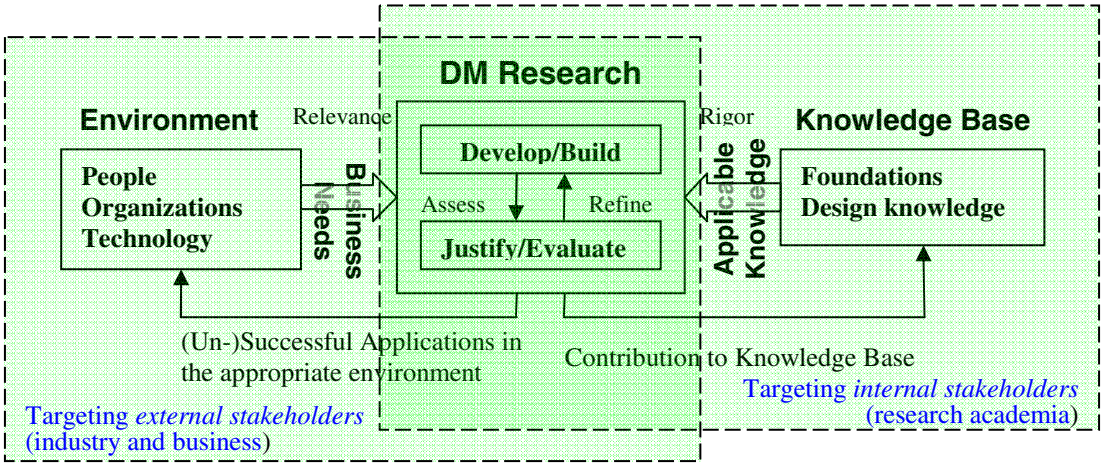

Figure 2. Rigor and relevance aspects of DM research and its "marketing"

external and internal stakeholders, as IS do. DM researchers themselves decide which ones of the stake-holders they consider as their customers. The narrowest scope is to think that only other DM researchers are our customers (as in reference [1] above).

Our opinion is that it is time to seriously consider broadening the scope of DM research to cover more topics related to the main external stakeholders too, i.e. to raise some external stakeholders as customers of our research results.

3. Relevance of DM research for stakeholders

Cresswell [5, p.2] writes "One rather critical distinction is between *relevance to* and *serves the interests of* or *is value to*" when considering the relevance of IS research. Hirschheim and Klein [13, p.249] keeps this distinction as a starting point in exploring the meaning of relevancy. According to them [13, pp.249-250] "Creswell points out that some research could be rather critical of practice and could undermine a stakeholder's interest, yet this would not make the research irrelevant".

They [13, p.249] continue that "as different stakeholder groups tend to possess conflicting interests arising from different value systems, IS relevancy depends on value judgments that should be made explicit while not engendering opportunities to learn from the interaction with any stakeholders that are willing to open themselves to IS researchers." They continue considering relevance through so called *disconnects* that represent differences of expectations between IS research and stakeholders.

In discussions concerning internal stakeholders the relevance for academics from other departments is considered to be at least as important as the relevance for external stake-holders [13, p.259], because the other academics "control the advancement of IS researchers and the field as a whole more than anything else". They see that these other communities may have entirely different sets of expectations to IS, even if they share the applied focus of IS.

The most important internal stakeholders of DM research are researchers from the same area. For them the rigor aspect of research is dominating the main criteria for relevance of research, too. Nowadays widespread utilization of IT within diverse industries (manufacturing, health care, education, etc) has also raised the interest of other academics to the results of IT-related research and also DM research. This means that we encounter a growing pull as DM researchers to make rigor research that at the same time produce useful results for researchers of other areas.

Hirschheim and Klein [13, pp.250-253] have recognized that both the business community and the academic community have not managed to justify their expectations about IS research. They blame the IS research community in having done a very poor job of communicating in a not very convinced way, if the IS researchers truly believe that their theories are relevant for practitioners. On the other hand "the view of IS held by IS-practitioners is at best only partially supported by some theories that guide IS research" [13, p.253] and the view of non-IS practitioners is still "even more at odds". As a way to start to solve these communication problems they suggest increasing the amount of research directed at understanding both the IS and non-IS practitioners and having discourse with them about realistic expectations with regard to IS. If this is still the common situation when IS researchers have had co-operation with practitioners over several decades, then what is the situation with the mutual understanding in DM research?

When considering various ways to address industry in IS research, Chiasson and Davidson [4] also outlined a range of strategies for incorporating industry into IS research. We adopt their reasoning to the context of DM

research. Indeed, DM researchers rarely take industry into consideration while conducting their often rigorous research activities. Although some exceptions can be found, as it is with some issues in active learning and cost-sensitive learning areas, which address some utility issues, such as cost-effective acquisition of information for the training data; consideration of costs and benefits associated with using the learned knowledge and how these costs and benefits should be factored into the DM process. Taking the needs of industry seriously into account is still in its infancy in the DM research area even when the industry context is important to the meaning, design, use and structure of a DM artifact. The situation is even more complicated because the outputs vary significantly with different branches of industry, affecting the meaning and measurement of utility and performance (especially e.g. with not-for-profit industries where the outputs are often complex).

Lin in Wu *et al.* [18] notices that a new successful industry (such 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 present there are several dozens of small-scale DM systems. This fact according to Lin indicates that we are still in the 3rd phase in the DM area. However, we believe that DM is going towards the next levels, and therefore the study of the DM development and DM use processes is equally important as the study of the technological aspects, and *such* research activities are likely to emerge within DM research community or outside it.

This is supported also by the recently established workshops and conference tracks, where applications of DM in industry/business and consideration of utility, associated risks and costs are encouraged.

By saying the above we are not arguing that industry/relevance/utility should be considered in every DM research project, though our analysis encourages the DM research community to take more relevance-oriented aspects of external stakeholders into its research agenda and thus considering them as their customers.

4. Beyond DM artifact

Until the mid-nineties DM required considerable specialized knowledge and was mainly restricted to users with substantial background in statistics, pattern recognition, databases and other related fields.

Dunkel *et al.* [10] concluded ten years ago that there is a need and opportunity for computing systems research and development, but still to the best of our knowledge there are no significant research papers published in this direction in the DM area.

Customer Relationship Management (CRM) software played a significant role in popularizing DM among corporate users. Availability of various DM algorithms, incorporation of DM modules by DB vendors in their solutions and emergence of open standard for accessing DM functionality from other applications also beneficially affected the attitude towards DM as a business function can provide a strategic advantage in developing, defining and deploying competitive business strategies.

However, it is still poorly recognized within the DM research community that it is essential to make research related to the development processes and use processes of DM systems, considering impacts and the essential factors that affect the impacts. We refer an interested reader to [16] where we adapted the IS success model of DeLone and McLean [9], Hevner's *et al.* [12] view on the behavioural-science paradigm and the design-science paradigm within the research in the IS discipline, and Nunamaker's *et al.* [14] view on system development as a multi-methodological IS research cycle to the DM area.

DM was earlier commonly considered as a separate part of the knowledge discovery process and this gave natural background to concentrate only on technological aspects behind the DM artifact, such as machine learning algorithms used in DM. But actually all the stages of the knowledge discovery process impact the utility of the knowledge derived from the data. The influence on the final utility of all the steps including acquiring data, extracting a model, and applying the acquired knowledge must also be considered. Similarly, utility considerations also impact the assessment of the decisions made based on the learned knowledge.

For us DM is inseparably included as an essential part of the knowledge discovery process, and we think that a more holistic view is needed to DM research. If this is accepted, the DM researchers have to take under investigation also the utility-related topics. Simple assessment measures like predictive accuracy have to give way to economic utility measures, such as profitability and return on investment.

We consider DM as a fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data [18]. DM researchers if interested to target their external stakeholders should recognize the major "marketing" challenges they are facing.

Some researchers might think that their task is only to continue doing a 'high-quality' (that often means rigorous but not necessary relevant) research, and an external stakeholder would easily find their results, and either apply them directly to address their well-understood need or in the worst case would be able to clearly formulate their need and ask to adapt research outputs accordingly. We might imagine that such scenario works when the link between DM research and its external stakeholders is well-established and full-scale DM systems are widely

(and successfully) adopted. Because this is not the current situation, then DM research has to 'market' the outputs of their research to the external stakeholders, i.e. to their potential customers.

DM researchers, if interested to target their external stakeholders, should recognize the major "marketing" challenges they are facing. They need to understand the needs of business and they need to communicate with their business actors to be able to adjust their expectations concerning DM possibilities to a realistic level.

We apply the framework of Smith [17] introduced for marketing knowledge management in an organization to the context of DM research. Marketing can be seen as a process that involves five major steps [17]: (1) defining *the type of need* that can be already present, latent, or absent (i.e. not recognized by a customer); (2) ensuring that the output of DM research meets the customer's need (so-called *brand awareness*); (3) stressing to a favorable attitude towards a brand (so-called *brand attitude*); (4) assisting the target customer to take an action using DM artifact (so-called *brand purchase intention*); and finally (5) facilitating purchase.

Marketing advices to recognize basic, enabling and strategic needs, define what are the current needs and focus on them (ensuring that lower level needs are continue to be met). It has been recognized that many failures of IT/IS were due to development of too generic capabilities which did not add business value [17]. DM was not an exception with such experiences. Consequently, a negative attitude towards DM as the result of problems with the past history may make it difficult to convince our external stakeholders to invest their (often limited) resources in DM rather than other options.

5. Concluding remarks

The current situation with DM research, namely its focus on rigor in research without taking many relevance (and utility) aspects seriously into consideration motivated us to consider what inspirations might come up looking at some articles related to the discussions of relevance going on in the IS discipline. In this paper we raised the question "who is our customer?" as DM researchers and discussed related topics starting from the stakeholders of DM research. We divided the major stakeholders of DM research into internal and external ones and considered some aspects of the relevance of DM research from their point of view. Our main aim is to raise under discussion what research topics are "acceptable" for DM researchers, i.e. who are our customers. If a more holistic view of DM research is selected then the DM research community needs to pay more attention to both the needs of larger group of customers and marketing the research results in a way that supports realistic expectations of them.

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

6. References

1. Alter S. "Sidestepping the IT Artifact, Scrapping the IS Silo, and Laying Claim to "Systems in Organizations", *Communications of the Association for Information Systems Vol. 12, Article 30,* 2003.
2. 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.
3. Bhattecherjee A., "Understanding and Evaluating Relevance in IS Research", *Communications of the Association for Information Systems Vol. 6, Article6,* 2001.
4. Chiasson M., Davidson E. "Taking industry seriously in Information Systems research", MIS Quarterly 29(4), Dec. 2005, pp. 591 – 605.
5. Cresswell A., "Thoughts on Relevance of IS Research", *Communications of the Association for Information Systems Vol. 6, Article 9,* 2001.
6. CRISP-DM: 1.0 *Step-by-step DM guide*, SPSS Inc.
7. Davenport T.H., Markus M.L. "Rigor vs. Relevance Revisited: Response to Benbasat and Zmud" ", *MIS Quarterly 23(2),* 1999, pp. 19 – 23.
8. 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.
9. 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
10. Dunkel B., Soparkar N., Szaro J., Uthurusamy R. "Systems for KDD: From concepts to practice", *Future Generation Computer Systems 13(2)*, 1997, pp. 231-242
11. Fayyad U. "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert 11(5)*, 1996, pp.20-25
12. Hevner A., March S., Park J., Ram S. "Decision Science in Information Systems Research", *MIS Quarterly 26*(1), 2004, pp. 75-105.
13. Hirschheim R., Klein H.K., "Crisis in the IS Field? A Critical Reflection on the State of the Discipline", *Journal of the Association for Information Systems 4(10),* 2003, pp. 237-293.
14. Nunamaker W., Chen M., Purdin T. "Systems development in information systems research", *Journal of Management Information Systems 7(3)*, 1990-91, 89-106.
15. Pechenizkiy M., Puuronen S., Tsymbal A. "Does the relevance of data mining research matter?" (to appear) *Foundations of Data Mining*, Springer, 2006.
16. Pechenizkiy M., Puuronen S., Tsymbal A. "Competitive advantage from Data Mining: Lessons learnt in the Information Systems field", In: *IEEE Workshop Proc. of DEXA'05, 1st Int. Workshop on Philosophies and Methodologies for Knowledge Discovery PMKD'05*, IEEE CS Press, 2005, pp. 733-737.
17. Smith H. "Developments in practice XIV: Marketing KM to the organization", *Journal of the Association for Information Systems 14*, 2004, pp. 513 – 525.
18. 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.