Towards more relevance-oriented data mining research

Mykola Pechenizkiy[a,*], Seppo Puuronen[b] and Alexey Tsymbal[c]
[a]*Department of Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[b]*Department of Computer Science and Information Systems, University of Jyv äskylä, P.O. Box 35, Jyväskylä 40351, Finland*
[c]*Siemens AG, CT SE SCR 2, Erlangen, Germany*

Abstract. Data mining (DM) research has successfully developed advanced DM techniques and algorithms over the last few decades, and many organisations have great expectations to take more benefit of their data warehouses in decision making. Currently, the strong focus of most DM-researchers is still only on technology-oriented topics. Commonly the DM research has several stakeholders, the major of which can be divided into internal and external ones each having their own point of view, and which are at least partly conflicting. The most important internal groups of stakeholders are the DM research community and academics in other disciplines. The most important external stakeholder groups are managers and domain experts who have their own utility-based interests to DM and DM research results. In this paper we discuss these practice-oriented points of view towards DM research and suggest broader discussions inside the DM research community about who should do that kind of research. We bring in the discussion several topics developed in the information systems (IS) discipline and show some similarities between IS and DM systems. DM systems have also their own peculiarities and we conclude that researchers who take into account human and organisational aspects related to DM systems need to have also some understanding about DM. This makes us suggest that the research area inside the DM community should be made broader than the current heavily technology-oriented one.

Keywords: Rigor vs. relevance in research, utility-based data mining, data mining stakeholders

1. Introduction

Data mining (DM) and knowledge discovery help to accumulate and process data and make use of it [16]. The two disciplines 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.

Technical aspects of DM have received good amount of rigorous 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 organisations 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 organisation, too. Currently the maturation of

[*] Corresponding author. E-mail: m.pechenizkiy@tue.nl.

DM-supporting processes which would take into account human and organisational aspects is still living its childhood.

There has been quite much research dedicated to DM frameworks. We have presented a comprehensive review of existing DM frameworks grouping them into three categories: theory-oriented, process-oriented, and foundation-oriented frameworks [26].

The theory-oriented frameworks are based mainly on one of the four following paradigms: 1) one of *the three statistical paradigms*: statistical experiment paradigm, statistical learning from empirical process paradigm, and structural data analysis paradigm, 2) *the data compression paradigm* where the dataset is compressed by finding some structure or knowledge within 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, evaluation, interpretation, and validation. CRISP-DM [5] is maybe the best example of a methodology for the DM artifact production process.

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. However, there are also strong opinions supporting diversity seeing an umbrella-framework as more reasonable. A similar discussion has been going on quite a while about the core of Information System (IS) research (see for example two recent works [1,2]).

Different theory-based 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 (as in [23]). Some recent work in this area has emerged within the so-called utility-based data mining (UBDM) research [34,37].

In this paper we focus on considering *the stakeholders of DM research* and their utility expectations with respect to DM research results. Corresponding considerations have earlier been taking place in other research areas, as in the IS discipline that we refer to. In the IS community an IS is often considered (in the traditional IS research framework [10]) in its organisational environment that is surrounded by an external environment.

We have suggested in [27] that user and organisation related research results and organisational settings used in the IS discipline include essential points of view which might be reasonable to take into account in developing DM research towards practically more relevant directions in domain areas where human and organisational aspects matter. As IS research, also DM research has several stakeholders, the major of which can be divided into internal and external, each having their own and commonly conflicting goals. Currently, DM researchers rarely take industry (the most important external stakeholder) into account while conducting their often technology oriented research activities.

So far in the DM community there are 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 (Fig. 1).

This even holds in the industry context where meaning, design, use, and structure of a DM artifact is an important topic. Situation is still more complicated because outputs vary significantly by industry affecting the meaning and measurement of utility and performance.

Although, recent development in cost-sensitive learning and active learning has started to consider some of the economic utility factors (like the cost of data, cost of measurement, cost of class label and so forth), yet many other utility factors are left outside the main directions of the emerging utility-based DM research (UBDM).

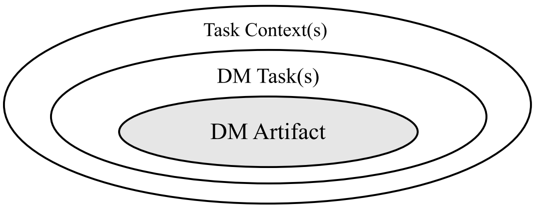

Fig. 1. DM artifact (adapted from [2]).

For us DM is inseparably included as an essential part of the knowledge discovery process and we see that the more holistic view of DM research is needed. If we as DM researchers want to participate in this kind of research efforts then we need to accept to take under investigation also utility-related topics. Simple assessment measures in DM like predictive accuracy have to give way to economic utility measures, such as profitability and return on investment. But on the other hand data mining systems (DMS) have their own peculiarities, which should be taken into account also in the holistic DMS research. We have introduced a generic framework for utility-based DM research in [29].

In this paper we motivate DM researchers to consider the possibility to take utility aspects into account from a broader perspective than is usually done in the UBDM context nowadays, and consider different groups of stakeholders of DM research and their corresponding (possibly conflicting) interests in utility considerations.

The rest of the paper is organised as follows. In Section 2 we introduce DM as a process and briefly review and summarise recent research directions in UBDM, which were reflected at UBDM-05 and UBDM-06 workshops [34,37], and some other relevant publications. In Section 3 we discuss the stakeholders of DM research, and motivate broadening the utility concept in the context of UBDM, emphasising the importance of business understanding and necessity to analyse the process of using/applying the developed DM artifact in the real context. In Section 4 we continue analysis of DM research stakeholders considering the relevance aspect of research from the different stakeholders' point of view. In Section 5 we address the issues of DM artifact development and use, and a broader vision beyond artifacts. We conclude briefly with some remarks and discussion of further research in Section 6.

2. DM processes: A relevance-oriented perspective

Fayyad in [16] defines knowledge discovery from databases (KDD) as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Before focusing on discussion of the DM process, we would like to make a note that this definition is pretty capacious; it gives an idea what is the goal of DM and in fact it is cited in many DM related papers in introductory sections. However, in many cases those papers have nothing to do with novelty, interestingness, potential usefulness, and validity of patterns which were discovered or could be discovered using DM techniques proposed later in those papers.

The DM process comprises many steps, which involve data selection, data preprocessing, data transformation, search for patterns, and interpretation and evaluation of patterns [16]. The steps start with the raw data and finish with the extracted knowledge, which was acquired as a result of the whole DM/KDD [1] process. The set of DM tasks used to extract and verify patterns in data form the core of the process. Most

[1] We would like to clarify that according to Fayyad's definition, and some other research literature, DM is commonly referred

current DM/KDD research is dedicated to the pattern mining algorithms and descriptive and predictive modelling of data.

The life cycle of a DM project according to the CRISP-DM model consists of six phases (though the sequence of the phases is not strict and moving back and forth between different phases normally happens) [5].

For us it is natural to define utility as an overall measure of benefit (e.g. economic) and thus the utilities connected with the entire DM process. In the rest of this section we review UBDM research directions and summarise the steps of the CRISP-DM process which are taken into account.

Considerations of costs and benefits are common for all managerial decisions in an organisation. Consequently, the quality of a DM artifact and its output must be evaluated considering its ability to enhance the quality of the resulting decision. Most of early works in predictive DM did not address the different practical issues related to data preparation, model induction, and its evaluation and application. Cost-sensitive learning [33] research emerged initially in DM as an effort to reflect the relevance of incorporating the costs resulting from a decision (based on prediction of DM model). Many application areas of DM suggested that e.g. for classification, the costs to predict class membership of instances accurately are proportional to the amount of accurately predicted instances yet needed to account for the asymmetric costs associated with true versus false prediction of positives and negatives. The knowledge of this asymmetry can be used to guide the parameterisation of a classifier and selection of the most appropriate one (e.g. MetaCost [13] or cost-sensitive boosting [15]). This leads to the development of robust evaluation techniques like the ROC convex hull method [28] or the area under the ROC curve (AUC) [4] which can be utilised when considering the business problem and managerial objectives.

As DM is commonly referred to as secondary data analysis, it is often assumed that a fixed amount of training data (being collected for some other purposes) is available for the current goal of knowledge discovery. Consequently, it is assumed by many developers of DM techniques that data is given and there are no costs associated with the availability of the data. However, sooner or later it becomes evident that availability of data for analysis (and especially the availability of labeled data for supervised learning) affects the economic utility of acquiring training data (or labelling unlabeled data) and therefore, should be considered as the costs of building a model, and applying the model.

Thus, e.g. in medical diagnostics the general problem can be formulated as follows: given the costs of tests and the total fixed budget, decide which tests to run on with which patients to obtain additional information needed to produce an effective classifier (assuming that no or little training data is available initially) [18]. Then, cost consideration includes the costs associated with building the classifier and the costs (and benefits) associated with applying the classifier.

Evaluation of cost-sensitive learners was studied by Holte and Drummond in [22]; they introduced cost curves which enable easy visualisation of average performance (expected cost), operating range, confidence intervals on performance, and difference in performance and its significance.

Thus, it seems that at least most of the current research in UBDM inclines to cost-sensitive learning and active learning paradigms from the machine learning perspective, and refers to total utility as derived from the following DM related processes:

to as a particular phase of the entire process of turning raw data into valuable knowledge, and it includes the application of modeling and discovery algorithms. In industry, however, both knowledge discovery and DM terms are often used as synonyms to the entire process of producing valuable knowledge from data.

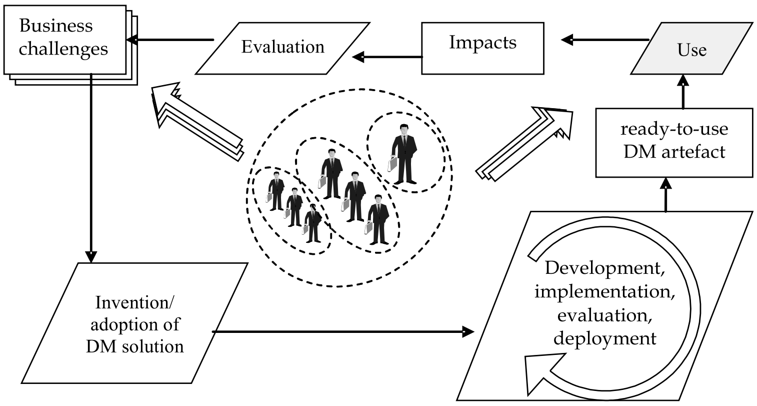
Business challenges
Evaluation
Impacts
Use
ready-to-use DM artefact
Development, implementation, evaluation, deployment
Invention/ adoption of DM solution

Fig. 2. DM Stakeholders view on utility considerations.

- *data preparation*[2] – costs of acquiring the data, including primarily the costs of 1) measuring an attribute value, 2) data labelling for supervised learning, 3) data records collection/purchase/retrieval, and 4) data cleaning and preprocessing;
- *data modelling and evaluation* – costs of search for patterns in the data, costs of misclassification, benefits of using the discovered patterns/models.[3]

Thus, deployment and impact estimation from the use of DM artifact processes, which we believe are at least not less important for addressing utility issues in UBDM as data preparation and data modelling processes, are currently almost completely ignored in DM research.

Even if UBDM researchers say that the goal of UBDM is to maximize the total benefit of using the mined knowledge minus the costs of acquiring and mining the data, yet it does not assume the thorough analysis of use processes and accounting for various benefits (and risks) associated with them. The mined knowledge is of utility to a person or an organisation if it contributes to reach a desired goal. Utility-based measures in itemset mining use the utilities of the patterns to reflect the user's goals. Yao et al. in [36] review utility-based measures for itemset mining and present a unified framework for incorporating several utility-based measures into the DM process by defining a unified utility function. Yet, it is always assumed that there is a (single type of) user and that the user is able to clearly formulate business challenges and to help to find an appropriate transformation of them into a set of DM tasks or simply pick up solutions among the available ones.

In the following section we consider different types (and organisational levels) of DM use and DM stakeholders emphasising the differences in utility considerations depending on the type of DM use or the type of DM stakeholder. We support our discussion with Fig. 2 having roots in CRISP-DM but emphasising the importance of *Use* process that can start as long as there exists a ready-to-use DM artifact (that is the result of the development, implementation, evaluation, and deployment of certain DM solution that address the recognised business challenge). We emphasise also that the *use* process is connected to a certain type(s) of DM stakeholders (yet potentially belonging to the same organisation), who may have different (and potentially competing) business challenges. The use of DM artifact leads

[2] Data preparation step can be associated not only with data preprocessing, data selection and data transformation processes, but also with data collection, data acquisition and/or data labeling.

[3] Currently this direction is limited to accounting for some economic benefits, known (or believed to be known) in advance without estimation of actual individual or organisational impact using the DM artifact.

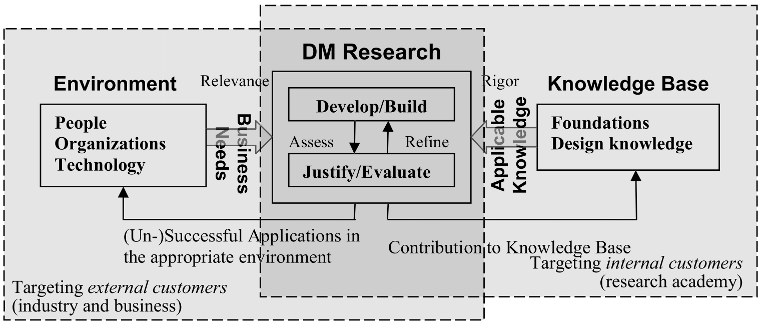
DM Research
Environment
Relevance
People
Organizations
Technology
Business Needs
Develop/Build
Assess
Refine
Justify/Evaluate
Rigor
Applicable Knowledge
Knowledge Base
Foundations
Design knowledge
(Un-)Successful Applications in the appropriate environment
Contribution to Knowledge Base
Targeting internal customers (research academy)
Targeting external customers (industry and business)

Fig. 3. Rigor and relevance aspects of DM research and DM stakeholders (adapted from [20]).

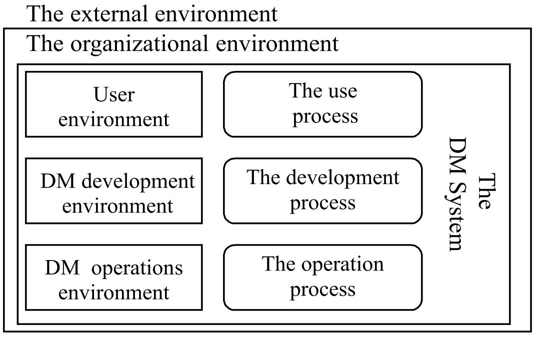

Fig. 4. A traditional IS's view applied to DM research (adapted from [10]).

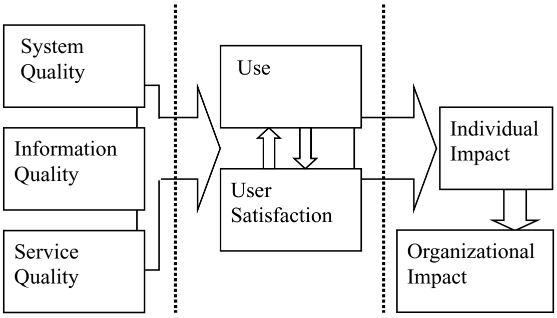

Fig. 5. Adapted from D&M IS Success Model [11] and updated D&M IS Success Model [12].

to certain individual and/or organisational impacts (illustrated later with Fig. 5), the evaluation of which in our opinion is exactly the main way of utility and relevance estimation.

3. Stakeholders of DM research

A *stakeholder* of DM research is a person or organisation that has a legitimate interest in DM research or its results. Following [30], 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 (Fig. 3).

Related to IS research, Hirschheim and Klein in [21] stress the need to recognise 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 IS 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 analyses of external stakeholders Hirschheim and Klein in [21] 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. in [20] have considered the design science setting from the relevance and rigor point of view. Figure 3 is a 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. in [20] support the idea that IS research at the best serve both the environment and the knowledge base.

Chiasson and Davidson in [6] considered various ways industry can be addressed in IS research; they also assessed how industry influences IS activities. They found that industry often provides important contexts and so-called *contextual spaces* to build a new theory and to refine or evaluate the boundaries of an existing theory.

There are opposite opinions also in the IS research area. For example Alter in [1] mentions referring to the rigor versus relevance discussions in [9] that "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 [35] 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 external and internal stakeholders, as IS do. DM researchers themselves decide which of them they take into account in their research settings. The narrowest scope is to think that only other DM researchers are considered (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.

4. Relevance of DM research for stakeholders

Cresswell in [8] 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 in [21] keep this distinction as a starting point in exploring the meaning of relevancy. According to them (ibid, 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 continue (ibid, 249) 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 its stakeholders.

Indeed, the utility considerations can be regarded to be at least partly different for different groups of DM research stakeholders.

If we imagine utility consideration by top-management in some organisation (i.e. external stakeholder, who can, potentially, adopt a DM artifact or results produced by it for managerial decision making) when deciding of DM project selection, then the traditional investments appraisal technique can be Parker et al.’s [25] information economics considering two domains: business and technology domain. The business domain includes the following factors: 1) return on investment, 2) strategic match, 3) competitive advantage, and 4) organisational risk. The technology domain includes the factors: 1) strategic architecture alignment, 2) definitional uncertainty risk, 3) technical uncertainty, and 4) technology infrastructure risk. Steward and Mohammed in [32] have in their article regrouped the factors of Parker et al. [25] under two main criteria: value and risk.

Utility consideration of a domain expert (i.e. another type of an external stakeholder group that uses DM artifact or results produced by it for decision support in daily operational decision making e.g. in diagnostics) includes sub-criteria that impact the overall utility of DM tool from his/her point of view such as: 1) satisfaction from use, 2) possible changes in responsibilities in decision making, 3) whether the tool is transparent in functionality and the results are easy to interpret, 4) whether training and support is likely to be needed (and provided), and finally, 5) what the overall impact from performance perspective of the use of the tool will be. We can see that these utility considerations (in fact, potentially related to the same DM artifact) differ quite a lot from the ones of the previous group of stakeholders.

In discussions concerning internal stakeholders of IS research the relevance for academics from other departments is considered to be at least as important as the relevance for external stake-holders, because the other academics “control the advancement of IS researchers and the field as a whole more than anything else” [21]. 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 of course researchers in the same area. For them the rigor aspect of research is dominating and the main criteria for relevance of research, too. Nowadays widespread utilisation 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 DM research, too. This means that we encounter a growing pull as DM researchers to make rigor research that at the same time produces useful results for researchers of other areas.

Hirschheim and Klein in [21] have recognised 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” [21] 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 realistic to expect about the situation with the mutual understanding in DM research?

When considering various ways to address industry in IS research, Chiasson and Davidson in [6] 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 [35] 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 the 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.

5. Beyond DM artifact

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

Dunkel et al. in [14] 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 popularising 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 which provides a strategic advantage in developing, defining, and deploying competitive business strategies.

However, it is still poorly recognised 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.

Piatetsky-Shapiro in [35] gives a good example that characterises 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".

In [27] we suggested adopting the traditional IS research framework [10] and, analogously, we considered DMS as a system having organisational and external environments and including a user environment, a DM development environment, and a DM operations environment (Fig. 4).

Indeed, DM is a fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data [35]. It is essential to make research related to the use processes of DMSs, considering impacts and the essential factors that affect the impacts. We introduced in [27] to DM community an adapted IS Success Model [11] and updated D&M IS Success Model [12] that are very well-known in the IS discipline (Fig. 5).

In IS, success factors research has a long tradition. A similar approach is needed with DMS to recognise the key factors of successful use and impact of DMS both at individual and organisational levels. The first efforts in that direction are the ones presented in the DM Review magazine [7,19] referred below and those should be followed by research-based reports.

Coppock in [7] analysed, in a way, the failure factors of DM-involved projects. In his opinion they often have nothing to do with the skill of the modeller 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 organisation are not less important than the traditionally emphasised technological job of turning data into insights.

Hermiz in [19] 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) recognising that DM is a process with many components and dependencies – the entire project cannot be "managed" in the traditional sense of the business word, and 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 [35] 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 in [17] 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 summarise or discover the main success factors and the research should be continued.

We refer an interested reader to [25], where we discuss applying an adapted Hevner et al.'s [20] view on the behavioural-science paradigm and the design-science paradigm within the research in the IS discipline, and Nunamaker et al.'s [24] 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 final utility is influenced by all the steps including acquiring data, extracting a model, and applying the acquired knowledge. 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 in DM research. If this is accepted, then 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.

If the DM community considers DM as a fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data then 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 at least would be able to clearly formulate their needs and ask to adapt research outputs accordingly. We might imagine that such a 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 its research to the external stakeholders. Researchers 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 in the DM research context the framework of Smith [31] introduced for marketing knowledge management in an organisation. Marketing can be seen as a process that involves five major steps [31]: 1) defining *the type of need* that can be already present, latent, or absent (i.e. not recognised by a customer), 2) ensuring that the output of DM research meets the customer's need (so-called *brand awareness*), 3) stressing to a favourable 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 recognise 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 recognised that many failures of IT/IS were due to development of too generic capabilities which did not add business value [31]. DM is not an exception having also 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.

6. Concluding remarks and future work

DM has lately been loaded with strong expectations to help organisations and individuals get more utility from their databases and data warehouses. These expectations are more based on the fine rigor research results achieved with technical aspects of DM methods and algorithms than the vast amount of practical success stories. The time when DM research has to answer also the practical expectations is fast approaching. Who should care about research of users' (both individuals' and organisations') goals and success factors when they install and use DM (system) parts in their ensembles of information system functionalities? Do the researchers of these topics need some amount of DM knowledge beside the information system one?

The goal of this paper is to raise these broader utility based DM questions under discussion of researchers in the DM area. We first made a short review of the DM research that has taken utility into account. Then, 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. After that we took a closer look at practice oriented normative advices included into the CRISP-DM model to motivate the use and user orientation more broadly in the utility considerations. We discussed briefly the well-known success model and the design science approach applied in the IS discipline. We considered more closely the stakeholders of DM research especially from their utilities point of view in a new revised framework that reflects DM related processes.

From a wider perspective we believe that building of knowledge networks across the field boundaries (DM and IS) would benefit the DM field in terms of better understanding and addressing such important issues as DM success, DM costs, DM risks, DM life cycles, methods for analysing DM systems, organising and codifying knowledge about DM systems in organisations, and maximising the value of DM research.

Our main aim is to raise under discussion what research topics are "acceptable" for DM researchers. 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.

In our future work we plan to focus on meta-analysis of the DM research tracing its development, and to produce categorisations based on theory/practice orientation of examined DM research, the use of different kinds of research methods, and other criteria.

We plan to estimate approximately the proportions of published work in different directions and different types of DM research through literature review and papers categorisation according to predefined classification criteria. Beside the analysis of the relevant literature from the top international data-mining related journals and international conference proceedings, we plan to collect and analyse the editorial policies of these top international journals and conferences. This will result in better understanding of the major findings and trends in the DM research area. We expect that this will also help us to highlight the existing disbalance in the area, and suggest the ways of improving the situation.

Acknowledgements

This research was partly supported by the Academy of Finland. We are thankful to the reviewers and to the guest editors of this special issue for their valuable comments and constructive criticism.

References

[1] S. Alter, Sidestepping the IT Artifact, Scrapping the IS Silo, and Laying Claim to "Systems in Organizations", *Communications of the Association for Information Systems* **12** (2003), Article 30.

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

[3] A. Bhattecherjee, Understanding and Evaluating Relevance in IS Research, *Communications of the Association for Information Systems* **6** (2001), Article 6.

[4] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition* **30** (1997), 1145–1159.

[5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, *CRISP-DM 1.0 – Step-by-step data mining guide,* The CRISP-DM Consortium / SPSS Inc. (2000), Available from http://www.crisp-dm.org.

[6] M. Chiasson and E. Davidson, Taking industry seriously in Information Systems research, *MIS Quarterly* **29**(4) (2005), 591–605.

[7] D.S. Coppock, Data Mining and Modelling: So You have a Model, Now What? *DM Review Magazine* (2003).

[8] A. Cresswell, Thoughts on Relevance of IS Research, *Communications of the Association for Information Systems* **6** (2001), Article 9.

[9] T.H. Davenport and M.L. Markus, Rigor vs. Relevance Revisited: Response to Benbasat and Zmud, *MIS Quarterly* **23**(2) (1999), 19–23.

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

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

[12] W. DeLone and E.R. McLean, The DeLone and McLean Model of Information Systems Success: A Ten-Year Update, *Journal of MIS* **19**(4) (2003), 9–30.

[13] P. Domingos, MetaCost: a general method for making classifiers cost-sensitive, in: *Proc. of the 5th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining,* San Diego, CA, USA, 1999, 155–164.

[14] B. Dunkel, N. Soparkar, J. Szaro and R. Uthurusamy, Systems for KDD: From concepts to practice, *Future Generation Computer Systems* **13**(2) (1997), 231–242.

[15] W. Fan, S.J. Stolfo, J. Zhang and P.K. Chan, AdaCost: Misclassification Cost-Sensitive Boosting, in: *Proc. of the 16th Intern. Conf. on Machine Learning,* Bled, Slovenia, 1999.

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

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

[18] R. Greiner, Budgeted Learning of Probabilistic Classifiers, in: *Proc. of the Workshop on Utility-Based Data Mining* (*UBDM'06*), Invited Talk, 2006.

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

[20] A. Hevner, S. March, J. Park and S. Ram, Design Science in Information Systems Research, *MIS Quarterly* **26**(1) (2004), 75–105.

[21] R. Hirschheim and H.K. Klein, Crisis in the IS Field? A Critical Reflection on the State of the Discipline, *Journal of the Association for Information Systems* **4**(10) (2003), 237–293.

[22] R.C. Holte and C. Drummond, Cost-sensitive classifier evaluation. In *Proceedings of the 1st Int. Workshop on Utility-Based Data Mining,* UBDM '05, ACM Press, New York, 2005, 3–9.

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

[24] W. Nunamaker, M. Chen and T. Purdin, Systems development in information systems research, *Journal of Management Information Systems* **7**(3) (1990–1991), 89–106.

[25] M. Parker, R. Benson and H. Trainer, *Information Economics: Linking Business Performance to Information Technology,* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[26] M. Pechenizkiy, S. Puuronen and A. Tsymbal, Does the relevance of data mining research matter? (to appear) In: *Foundations of Data Mining,* T.Y. Lin et al., eds, Springer, 2007.

[27] M. Pechenizkiy, S. Puuronen and A. Tsymbal, 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, 733–737.

[28] F. Provost and T. Fawcett, Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, in: *Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining,* Newport Beach, CA, USA, 1997.

[29] S. Puuronen and M. Pechenizkiy, *Towards the Generic Framework for Utility Considerations in Data Mining Research,* manuscript under review, 2007.

[30] S. Puuronen, M. Pechenizkiy and A. Tsymbal, Data Mining Researcher, Who is Your Customer? Some Issues Inspired by the Information Systems Field, in: *Proc. of the 17th international Conference on Database and Expert Systems Applications DEXA'06,* IEEE Computer Society, Washington, DC, 2005, 579–583.

[31] H. Smith, Developments in practice XIV: Marketing KM to the organization, *Journal of the Association for Information Systems* **14** (2004), 513–525.

[32] R. Steward and S. Mohammed, IT/IS Projects Selection Using Multi-criteria Utility Theory, *Logistics Information Management* **15**(4) (2002), 254–270.

[33] S. Viaene and G. Dedene, Cost-sensitive learning and decision making revisited, *European Journal of Operational Research* **166** (2004), 212–220.

[34] G. Weiss, M. Saar-Tsechansky and B. Zadrozny, *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining,* Chicago, Illinois, ACM Press, New York, NY, USA, 2005.

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

[36] H. Yao and H.J. Hamilton, Mining itemset utilities from transaction databases, *Data and Knowledge Engineering* **59**(3) (2006), 603–626.

[37] B. Zadrozny, G. Weiss and M. Saar-Tsechansky, *UBDM '06: Proceedings of the 2nd Int. Workshop on Utility-based data mining,* Chicago, Illinois, ACM Press, Philadelphia, Pennsylvania, USA, 2006.