Does Relevance Matter to Data Mining Research?

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

[1]Department of Computer Science and Information Systems, University of Jyväskylä, P.O.Box 35, FIN-40351, Jyväskylä, Finland, mpechen@cs.jyu.fi, sepi@cs.jyu.fi
[2] Department of Computer Science, Trinity College Dublin, Dublin 2, Ireland, Alexey.Tsymbal@cs.tcd.ie

Abstract

Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it. We review several existing frameworks for DM research that originate from different paradigms. These DM frameworks mainly address various DM algorithms for the different steps of the DM process. Recent research has shown that many real-world problems require integration of several DM algorithms from different paradigms in order to produce a better solution elevating the importance of practice-oriented aspects also in DM research. In this paper we strongly emphasize that DM research should also take into account the relevance of research, not only the rigor of it. Under relevance of research in general, we understand how good this research is in terms of the utility of its results. This chapter motivates development of such a new framework for DM research that would explicitly include the concept of relevance. We introduce the basic idea behind such framework and propose one sketch for the new framework for DM research based on results achieved in the information systems area having some tradition related to the relevance aspects of research.

1. Introduction

Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it [14]. 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 [41]. These processes include data cleaning, feature transformation, algorithm and parameter selection, and evaluation, interpretation and validation (Figure 1).

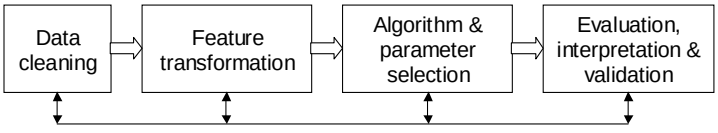

Fig. 1 - Data mining process (adapted from [41])

The idea of learning from data is far from being new. However, likely due to the developments in the database management field and due to the great 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 DM focuses on the development of new algorithms or improvement in the speed or the accuracy of the existing ones [30].

Relatively little has been published about theoretical frameworks of DM. A few theoretical approaches to DM were considered in [30]. A motivation for DM foundations development, and requirements for a theoretical DM framework were also considered in [30]: a theoretical framework should be simple and easy to apply; it should contribute to DM algorithms and DM systems development; it should be able to model typical DM tasks like clustering, classification and rule discovery; and it should recognize that DM is an iterative and interactive process, where a user has to be involved.

In this paper (in Section 2) we consider (1) several existing foundations-oriented frameworks for DM based on statistical, data compression, machine learning, philosophy of science, and database paradigms and (2) the most well-known process-oriented frameworks, including Fayyad [13], CRISP-DM [8], and Reinartz's [38] frameworks. We consider their advantages and limitations analyzing what these approaches are able to explain in the DM process and what they do not. We believe that a reader will notice that each one of the considered foundations-oriented DM frameworks is limited mainly to address one particular type of DM algorithms or describe certain view on the nature of DM. Process-oriented frameworks try to emphasize the issues of integration, iteration, and interactivity in DM. However, none of the frameworks stress the importance of *relevance* in DM research, i.e. they do not emphasize that relevant and applicable results from real world point of view will be achieved.

In empirical type of research, relevance usually appears to be associated with utility in practical applications. The so-called "richness of worldly realism" [31] associated with relevance is opposed to "tightness of control" [31] so that at the same level of knowledge they form an iso-epistemic curve representing the fundamental trade-off [24].

In design-science type of research relevance of research is often associated with the consideration of some business need(s), and related environment [20].

In this chapter we try to analyze whether relevance matters to DM research from both perspectives.

We need to acknowledge that some work has been done with regard to the study of *interestingness* of discovered patterns in the context of association rules mining (for example [39]). Yet, even within this particular area, it has been fairly noticed in [5] that there is no consensus on how the interestingness of discovered patterns should be measured, and that most of DM research avoids this thorny way reducing *interestingness* to *accuracy* and *comprehensibility*.

Disregarding the relevance issues, DM frameworks ignore also the issues of DM artifact development and DM artifact use. Here and in the following text by *DM artifact* we mean either "hard/technical" artifacts like DM model, DM technique or its instantiation, collection of DM techniques that are part of DM system or DM embedded solution, or "soft/social" artifacts like some organizational, operational, ethical and methodological rules that focus on different considerations of risks, costs, etc.

In Section 3 we first refer to the traditional information system (IS) framework presented in [9] that is widely known in the IS community and is a synthesis of many other frameworks considered before it. This framework takes into account

both the use and development aspects beside the technical ones in the IS area. Further we consider more detailed IS frameworks from the use and development perspectives.

In Section 4 we introduce our sketch for the new framework for DM research based on the material included in Sections 2 and 3. We strongly emphasize the relevance aspect of DM research, trying not to neglect the rigor. This means that beside the technological aspects also the organizational and human aspects should be equally taken into account. Thus, our framework for DM research suggests a new turning point for the whole DM research area.

We conclude briefly in Section 5 with a short summary and further research topics.

Some materials presented in this chapter are the results of our earlier work [35, 36, 37].

2. Review of some existing theoretical frameworks for DM

It this section we review basic existing foundations-oriented frameworks for DM based on different paradigms, originating from statistics, machine learning, databases, philosophy of science, and granular computing and the most well-known process-oriented frameworks, including Fayyad [13], CRISP-DM [8], and Reinartz's [38] frameworks. We present our conclusions for these groups of DM frameworks and then analyze the state of art in DM research in general.

2.1 Foundations(Theory)-oriented frameworks

Frameworks of this type are based mainly on one of the following paradigms: (1) the *statistical paradigms*; (2) *the data compression paradigm* – "compress the dataset by finding some structure or knowledge for it"; (3) *the machine learning paradigm* – "let the data suggest a model" that can be seen as a practical alternative to the statistical paradigms "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" [22]; and also *the inductive databases paradigm* – "locating interesting sentences from a given logic that are true in the database" [3].

2.1.1. The statistical paradigms

Generally, it is possible to consider the task of DM from the statistical point of view, emphasizing the fact that DM techniques are applied to larger datasets than it is commonly done in applied statistics [18]. Thus the analysis of appropriate statistical literature, where strong analytical background is accumulated, would solve most DM problems. Many DM tasks naturally may be formulated in the statistical terms, and many statistical contributions may be used in DM in a quite straightforward manner [17].

According to [6] there exist two basic statistical paradigms that are used in theoretical support for DM. The first paradigm is so-called "Statistical experiment". It can be seen from three perspectives: 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 is the “Statistical learning from empirical process” paradigm [40]. Generally, many DM tasks can be seen as the task of finding the underlying joint distribution of variables in the data. Good examples of this approach would be a Bayesian network or a hierarchical Bayesian model, which give a short and understandable representation of the joint distribution. DM tasks dealing with clustering and/or classification fit easily into this approach.

The second statistical paradigm is called “Structural data analysis” and can be associated with singular value decomposition methods, which are broadly used, for example, in text mining applications.

A deeper consideration of DM and statistics can be found in [15]. Here, we only want to point out that the volume of the data being analyzed and the different educational background of researchers are not the most important issues that constitute the difference between the areas. DM is an applied area of science and limitations in available computational resources is a big issue when applying results from traditional statistics to DM. An important point here is that the theoretical framework of statistics is not concerned much about data analysis as an iterative process that generally includes several steps. However, there are people (mainly with strong statistical background) who consider DM as a branch of statistics, because many DM tasks may be perfectly represented in terms of statistics.

2.1.2. The data compression paradigm

The data compression approach to DM can be stated in the following way: compress the dataset by finding some structure or knowledge within it, where knowledge is interpreted as a representation that allows coding the data using a fewer amount of bits. For example, the minimum description length (MDL) principle [32] can be used to select among different encodings accounting for both the complexity of a model and its predictive accuracy.

Machine learning practitioners have used the MDL principle in different interpretations to recommend that even when a hypothesis is not the most empirically successful among those available, it may be the one to be chosen if it is simple enough. The idea is in balancing between the consistency with training examples and the empirical adequacy by predictive success as it is, for example, with accurate decision tree construction. Bensusan [2] connects this to another methodological issue, namely that theories should not be *ad hoc*, that is they should not simply overfit all the examples used to build it. Simplicity is the remedy for being *ad hoc* both in the recommendations of the philosophy of science and in the practice of machine learning.

The data compression approach has also connections with the rather old Occam’s razor principle that was introduced in the 14th century. The most commonly used formulation of this principle in DM is "when you have two competing models which make exactly the same predictions, the one that is simpler is better".

Many (if not all) DM techniques can be viewed in terms of the data compression approach. For example, association rules and pruned decision trees can be viewed as ways of providing compression of parts of the data. Clustering can also be considered as a way of compressing the dataset. There is a connection

with the Bayesian theory for modeling the joint distribution – any compression scheme can be viewed as providing a distribution on the set of possible instances of the data.

2.1.3. The machine learning paradigm

The machine learning (ML) paradigm, "let the data suggest a model", can be seen as a practical alternative to the statistical paradigm "fit a model to the data". It is certainly reasonable in many situations to fit a small dataset to a parametric model based on a series of assumptions. However, for applications with large volumes of data under analysis the ML paradigm may be beneficial because of its flexibility with a nonparametric, assumption-free nature.

We would like to focus here on the constructive induction approach. Constructive induction is a learning process that consists of two intertwined phases, one of which is responsible for the construction of the "best" representation space and the second concerns generating hypotheses in the found space [33]. Constructive induction methods are classified into three categories: data-driven (information from the training examples is used), hypothesis-driven (information from the analysis of the form of intermediate hypothesis is used) and knowledge-driven (domain knowledge provided by experts is used) methods. Any kind of induction strategy (implying induction, abduction, analogies and other forms of non-truth preserving and non-monotonic inferences) may potentially be used. However, the focus here is usually on operating higher-level data-concepts and theoretical terms rather than pure data.

Many DM techniques that apply wrapper/filter approaches to combine feature selection, feature extraction, or feature construction processes (as means of dimensionality reduction and/or as means of search for better representation of the problem) and a classifier or other type of learning algorithm may be considered as constructive induction approaches.

2.1.4. The database paradigm

A database perspective on DM and knowledge discovery was introduced in [22]. The main postulate of their approach is: "there is no such thing as discovery, it is all in the power of the query language". That is, one can benefit from viewing common DM tasks not as the dynamic operations constructing the new pieces of information, but as operations finding unknown (i.e. not found so far) but existing parts of knowledge.

In [3] an inductive databases framework for the DM and knowledge discovery in databases (KDD) modeling was introduced. The basic idea here is that the data-mining task can be formulated as locating interesting sentences from a given logic that are true in the database. Then knowledge discovery from data can be viewed as querying the set of interesting sentences. Therefore the term "an inductive database" refers to such a type of databases that contains not only data but a theory about the data as well [3].

This approach has some logical connection to the idea of deductive databases, which contain normal database content and additionally a set of rules for deriving new facts from the facts already present in the database. This is a common inner data representation. For a database user, all the facts derivable from the rules are

presented, as they would have been actually stored there. In a similar way, there is no need to have all the rules that are true about the data stored in an inductive database. However, a user may imagine that all these rules are there, although in reality, the rules are constructed on demand. The description of an inductive database consists of a normal relational database structure with an additional structure for performing generalizations. It is possible to design a query language that works on inductive databases. Usually, the result of a query on an inductive database is an inductive database as well. Certainly, there might be a need to find a solution about what should be presented to a user and when to stop the recursive rule generation while querying. We refer an interested reader to [3] for details.

2.1.5. Granular-computing approach

Generally, granular computing is a broad term covering theories, methodologies, and techniques that operate with subsets, classes, and clusters (called granules) of a universe. Granular computing concept is widely used in computer science and mathematics. Recently, Zadeh [43] reviewed the concepts of fuzzy information granulation and considered it in the context of human reasoning and fuzzy logic. Lin [27] proposed to use the term ”granular computing” to label the computational theory of information granulation. In the same paper Lin introduces a view on DM as a ”reverse” engineering of database processing. While database processing organizes and stores data according to the given structure, DM is aimed at discovering the structure of stored data. Lin defines automated DM as *“a process of deriving interesting (to human) properties from the underlying mathematical structure of the stored bits and bytes”* [27]. Assuming that the underlying mathematical structure of a database relation is a set of binary relations or a granular structure, Lin considers DM as a processing of the granules or structure-granular computing. And then if there is no additional semantics, then the binary relations are equivalence relations and granular computing reduces to the rough set theory [27]. However, since in the DM process the goal is to derive also the properties of stored data, additional structures are imposed. To process these additional semantics, Lin introduces the notion of granular computing in DM context [28].

Yao and Yao [42] applied the granular computing approach to machine learning tasks focusing on covering and partitioning in the process of data mining and showed how the commonly used ID3 and PRISM algorithms can be extended with the granular computing approach.

2.1.6. The philosophy of science paradigm

The categorization of subjectivist and objectivist approaches [4] can be considered in the context of DM. The possibility to compare nominalistic and realistic ontological believes gives us an opportunity to consider data that is under analysis as descriptive facts or constitutive meanings. The analysis of voluntaristic as opposed to deterministic assumptions about the nature of every instance constituting the observed data directs our attitude and understanding of that data. One possibility is to view every instance and its state as determined by the context and/or a law. Another position consists in consideration of each instance as

autonomous and independent. An epistemological assumption about how a criterion to validate knowledge discovered (or a model that explains reality and allows making predictions) can be constructed may impact the selection of appropriate DM technique. From the positivistic point of view such a model-building process can be performed by searching for regularities and causal relationships between the constitutive constructs of a model. And anti-positivism suggests analyzing every individual observation trying to understand it and making an interpretation. Probably some of case-based reasoning approaches can be related to anti-positivism's vision of the reality.

An interesting difference in the views on reality can be found considering ideographic as opposed to nomothetic methodological disputes. The nomothetic school does not see the real world as a set of random happenings. And if so, there must be rules that describe some regularities. Thus, nomothetic sciences seek for establishing abstract (general) laws that describe indefinitely repeatable events and processes. On the contrary, the ideographic sciences are aimed to understand unique and non-recurrent events. They have connection to the ancient doctrine that "all is flux". If everything were always changing, then any generalization intending to be applied for two or more presumably comparable phenomena would never be true. And 'averages' of some measures (from the nomothetic way of thinking) usually is not able to represent the behaviour of a single event or entity.

2.1.7. Conclusion on the theory-oriented frameworks

The reductionist approach of viewing DM in terms of one of the theory-oriented frameworks has advantages in strong theoretical background, and easy-formulated problems. The statistics, data compression and constructive induction paradigms have relatively strong analytical background, as well as connections to the philosophy of science. In addition to the above frameworks there exists an interesting microeconomic view on DM [26], where a utility function is constructed and it is tried to be maximized. The DM tasks concerning processes like clustering, regression and classification fit easily into these approaches. Other small-scale yet valuable study related to analysis of interestingness measures of association rules is worth mentioning. Carvalho and Freitas [5], recognizing the potential gap in estimates of interestingness obtained with objective data-driven measures and true subjective evaluation performed by human, investigated the effectiveness of several data-driven rule interestingness measures by comparing them with the subjective real human interest.

One way or another, we can easily see the exploratory nature of the frameworks for DM. Different frameworks account for different DM tasks and allow preserving and presenting the background knowledge. However, what seems to be lacking in most theory-oriented approaches, are the ways for taking the iterative and interactive nature of the DM process into account [30], and a focus on the utility of DM.

2.2 Process-oriented frameworks

Frameworks of this type are known mainly because of works [13] and [8]. They view DM as a sequence of iterative processes that include data cleaning, feature

transformation, algorithm and parameter selection, and evaluation, interpretation and validation.

2.2.1 Fayyad's view on the knowledge discovery process

Fayyad [13, p.84] define KDD as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Before focusing on discussion of KDD as a process, we would like to make a note that this definition given by Fayyad is very capacious, it gives an idea what is the goal of KDD 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 proposed in the papers DM techniques.

KDD process comprises many steps, which involve data selection, data pre-processing, data transformation, DM (search for patterns), and interpretation and evaluation of patterns (Figure 2) [13]. The steps depicted start with the raw data and finish with the extracted knowledge, which was acquired as a result of the KDD process. The set of DM tasks used to extract and verify patterns in data is the core of the process. DM consists of applying data analysis and discovery algorithms for producing a particular enumeration of patterns (or models) over the data. Most of current KDD research is dedicated to the DM step. We would like to clarify that according to this scheme, and some other research literature, DM is commonly referred to as a particular phase of the entire process of turning raw data into valuable knowledge, and covers 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 getting valuable knowledge.

Nevertheless, this core process of search for potentially useful patterns typically takes only a small part (estimated at 15%-25%) of the effort of the overall KDD process. The additional steps of the KDD process, such as data preparation, data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results of mining, are also essential to derive useful knowledge from data.

In our opinion the main problem of the framework presented in Figure 2 is that all KDD activities are seen from "inside" of DM having nothing to do with the relevance of these activities to practice (business).

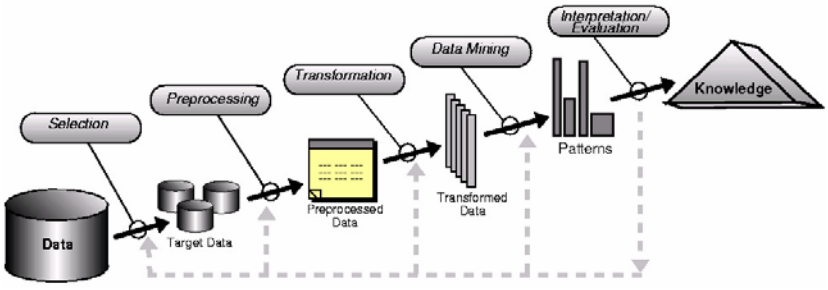

Fig. 2. Basic steps of the KDD process [13, p.85]

2.2.2 CRISP-DM

The life cycle of a DM project according to the CRISP-DM model (Figure 3) consists of six phases (though the sequence of the phases is not strict and moving back and forth between different phases normally happens) [8]. The arrows indicate the most important and frequent dependencies between phases. And the outer circle in the figure denotes the cyclic nature of DM – a DM process continues after a solution has been deployed. If some lessons are learnt during the process, some new and likely more focused business questions can be recognized and subsequently new DM processes will be launched.

We will not stop at any phase of CRISP-DM here since it has much overlapping with Fayyad's view and with the framework that is considered in the next subsection and discussed in more details. However, we would like to notice that the KDD process is put now in a way into some business environment that is represented by the *business understanding* and *deployment* blocks.

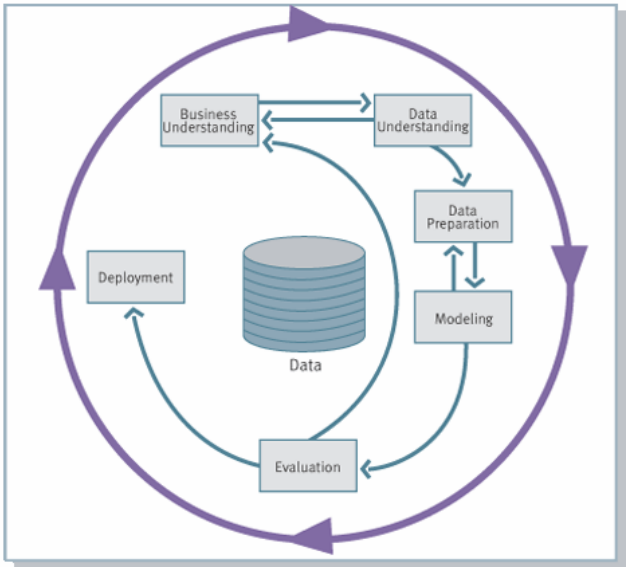

Fig. 3. CRoss Industry Standard Process for Data Mining [10]

2.2.3. Reunartz's view

Reinartz's framework [38] follows CRISP-DM with some modifications (Figure 4), introducing a data exploration phase and explicitly showing the accumulation of the experience achieved during the DM/KDD processes. The business-understanding phase is aimed to formulate business questions and translate them into DM goals. The data-understanding phase aims at analyzing and documenting the available data and knowledge sources in the business according to the formulated DM goals and providing initial characterization of data. The data preparation phase starts from target data selection that is often related to the problem of building and maintaining useful data warehouses. After selection, the target data is preprocessed in order to reduce the level of noise, preprocess the

missing information, reduce data, and remove obviously redundant features. The data exploration phase aims at providing the first insight into the data, evaluate the initial hypotheses, usually, by means of descriptive statistics and visualization techniques. The DM phase covers selection and application of DM techniques, initialization and further calibration of their parameters to optimal values. The discovered patterns that may include a summary of a subset of the data, statistical or predictive models of the data, and relationships among parts of the data are locally evaluated. The evaluation and interpretation phase aims at analyzing the discovered patterns, determining the patterns that can be considered as the new knowledge, and drawing conclusions about the whole discovery process as well. The deployment phase aims at transferring DM results that meet the success criteria into the business [38].

We think that the main problem with CRISP-DM and Reinartz's frameworks is that they assume that the DM artifact is ready to be applied and easy to be deployed and used. Therefore, the development and use processes are almost disregarded in these frameworks though being embedded in an implicit way. Consequently, it is hard to see what the most crucial success factors of a DM project are.

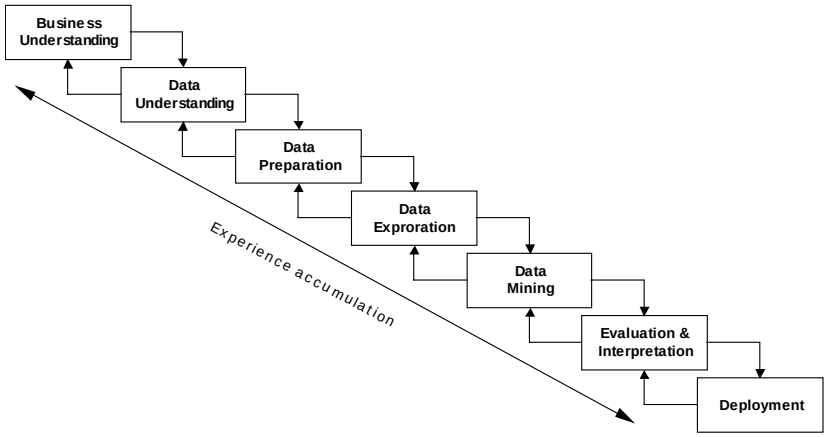

Fig. 4. Knowledge discovery process: from problem understanding to deployment (adapted from [38])

2.4 Conclusions on the considered frameworks

With respect to foundations-oriented frameworks, some DM researchers argue for the lack of an 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. And the approaches based on granular and rough computing present good examples of such attempts. 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.

SPSS whitepaper [8] states that "Unless there's a method, there's madness". It is accepted that just by pushing a button someone should not expect useful results to appear. An industry standard to DM projects CRISP-DM is a good initiative and a starting point directed towards the development of DM meta-artifact (methodology to produce DM artifacts). However, in our opinion it is just one guideline, which is in too general-level, that every DM developer follows with or without success to some extent. Process-oriented frameworks try to address the iterativeness and interactiveness of the DM process. However, the development process of DM artifact and use of that artifact are poorly emphasized.

Lin in Wu *et al.* [41] 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 the 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 at the 3rd phase with the DM area.

Further Lin in Wu *et al.* [41] 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 field many years ago, and it has resulted in introduction of several interesting IS research frameworks, some of which are discussed in the next section.

3. Information systems research frameworks

Information Systems (IS) are powerful instruments for organizational problem solving through formal information processing [29]. It is very common, especially in the US to use the term Management Information Systems (MIS) as a synonym for IS. From the first definitions of MIS in the first half of 1970s it has been developed as a discipline of its own having unique identity, core journals and conferences, and an official association with thousands members worldwide [1]. During the years different IS research frameworks have been defined and used. We represent in this chapter first the very traditional ones and then more resent

ones for the subareas of IS use and development.

3.1. The traditional information systems perspective

The traditional framework presented by Ives et al. [23] is widely known in the IS community. They used five earlier research models as a base when they developed their own (Figure 5). In their framework an IS is considered in an organizational environment that is further surrounded by an external environment. According to their 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.

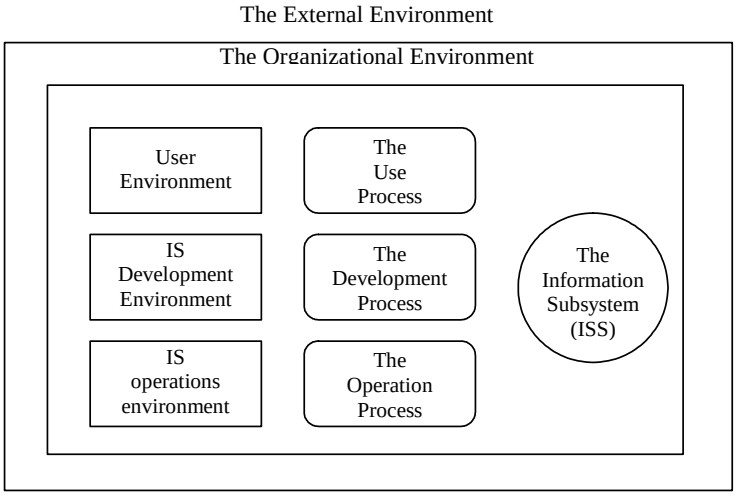

Fig. 5. A framework for IS research [23, p. 917]

The external environment [23, p. 916] ″includes legal, social, political, cultural, economic, educational, resource and industry/trade considerations″ and the organizational environment [23, p. 916] ″is marked by the organizational goals, tasks, structure, volatility, and management philosophy/style″. In their model the user environment is including and surrounding the primary users of the IS, the development environment consists of wide range of things from the technical (as IS development methods and techniques) to the organizational (as organization and management of IS development and maintenance) and human-oriented ones (as IS design personnel and their characteristics). The IS operations environment [23, p. 918] ″incorporates the resources necessary for IS operations. The major components include software, hardware, database, procedures/documentation, organization and management of IS operations, and the operations personnel″. However, in this paper, we focus on the user and IS development environments and the corresponding processes.

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 onto its organizational and external environments. Many research paradigms have been suggested and used in the IS discipline. Currently, Hevner *et al*. [20] suggest that two paradigms should be recognized 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" [20, p. 75]. This paradigm is naturally the most broadly applied in the use process related topics. They continue [20, p. 75] that "The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts". This second paradigm is the most natural in the IS development related topics where the new user and development environments are planned and experimented with. Some others as e.g. Iivari *et al.* [21] call the IS development process related research as a constructive type of research because it is based on the philosophical belief that development always involves creation of some new artifacts – conceptual (models, frameworks) or more technical artifacts (software implementations).

3.2. The IS success model

As one of IS user environment related models we represent in this section the IS success model developed by DeLone and McLean in 1992 [10]. They report in their ten-year update paper [11] that it has gained wide popularity with nearly 300 articles published in refereed journals referencing their original paper. Because this IS success model is so well known we picked it up as an example of user environment related IS research models. An adapted version of the model is presented in Figure 6 (It is very similar to the one in http://business.clemson.edu/ISE/).

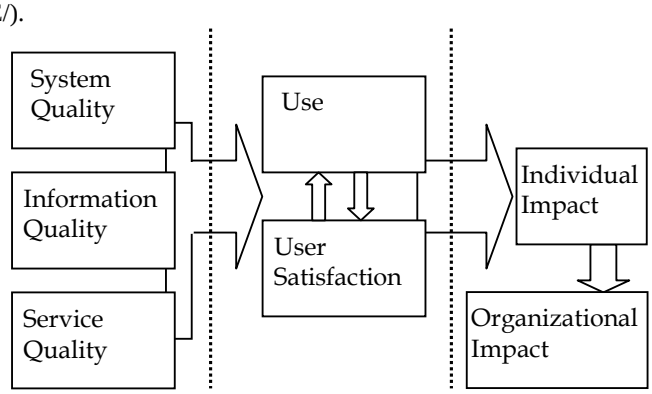

Fig. 6. Adapted from D&M IS Success Model [10, p. 87] and updated D&M IS Success Model [11, p. 24]

The original model was developed to "aid in the understanding of the possible causal interrelationships among the dimensions of success and to provide a more parsimonious exposition of the relationships". The investments into information systems are huge every year. Thus it is natural to try to evaluate the effectiveness

of those expenditures. The model raises information quality, service quality, and systems quality as key ingredients behind the user satisfaction and the use of IS. These have been found to have essential positive effect to individual impact leading to the organizational impact of information systems.

3.3. The IS development environment

The IS development environment is needed to develop and maintain the IS in use. Beside organizing and managing the development and maintenance processes these processes require several kinds of resources: not only the technical ones, as methods and techniques, but also human as motivated people with good enough education for the job. It is natural that in this compound human, organization, and technology complex there is a need to have diversified research methods. One such proposal that has been referred to quite often in the IS literature is the one represented below.

In [34] system development itself is considered as a central part of a multi-methodological information systems research cycle (Figure 7).

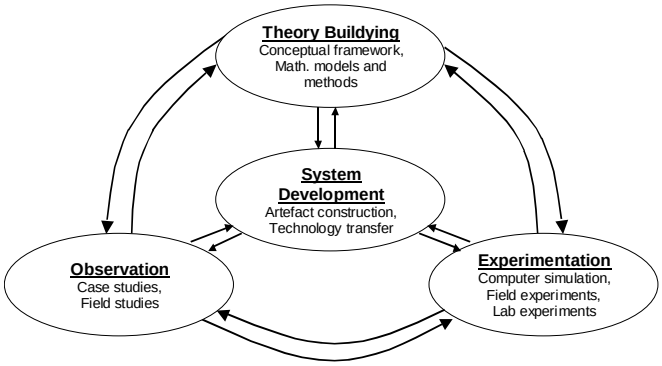

Fig. 7. A multimethodological approach to the construction of an artifact for DM (adapted from [34]).

Theory building involves 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 tested in the real world to show its validity, recognize its limitations and make refinements according to observations made during its application. According to reasoning research methods can be subdivided into basic and applied research, as naturally both are common for any large system development project [34]. A 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. Such a view presents the framework of IS as a complete, comprehensive and dynamic research process. It allows multiple perspectives and flexible choices of methods to be applied during different stages of the research process.

In fact, although Dunkel et al. [12] concluded that there is a need and opportunity for computing systems research and development in the context of DMS development, almost 9 years later, to the best of our knowledge there are no significant research papers published in this direction.

4. Our new research framework for DM research

It was mentioned in Section 2 that a new successful industry can follow five consecutive phases and that DM is presumably currently at the 3rd phase. The IS discipline on the other hand has during its 30+ year existence been able to develop to the 5th level. One of the key aspects helping the IS area development might have been that it has taken seriously into account human and organizational aspects beside the technological ones. This has raised its relevance and thus attracted more broad interests to support research in the IS area. We see raising the relevance of DM research as an essential aspect towards its more broad applicability, leading to new previously unknown research topics in the DM area.

In this section we suggest a new research framework which includes parts having similarities with the research frameworks applied in the IS discipline. Analogically with the IS research discussion in Section 3 we distinguish three environments for a DM system (DMS): the user, development, and operation environment but discuss in this paper only the first two. We start from these environments in Sections 4.1 and 4.2 and finish with presenting our new research framework for DM in Section 4.3.

4.1 The DMS user environment

Piatetsky-Shapiro in Wu *et al.* [41] gives a good example that characterizes the whole area of current DM research: “we see many papers proposing incremental refinements in association rules algorithms, but very few papers describing how the discovered association rules are used”. DM is fundamentally application-oriented area motivated by business and scientific needs to make sense of mountains of data [41]. A DMS is generally used to support or do some task(s) by human beings in an organizational environment (see Figure 8) both having their desires related to DMS. Further, the organization has its own environment that has its own interest related to DMS, for example that privacy of people is not violated.

A similar approach to that with IS is needed with DMS to recognize the key factors of successful use and impact of DMS both at the individual and organizational levels. Questions like (1) how the system is used, and also supported and evolved, and (2) how the system impacts and is impacted by the contexts in which it is embedded are important also in the DMS context. The first efforts in that direction are the ones presented in the DM Review magazine [7, 19], referred below. We believe that such efforts should be encouraged in DM research and followed by research-based reports.

Coppock [7] analyzed, in a way, the failure factors of DM-related projects. In his opinion they have nothing to do with the skill of the modeler or the quality of data. But those do include these four: (1) persons in charge of the project did not *formulate actionable insights*, (2) the sponsors of the work did not *communicate the insights* derived to key constituents, (3) the results *don't agree with*

institutional truths, and (4) the project never had a *sponsor and champion*. The main conclusion of Coppock's analysis is that, similar to an IS, the leadership, communication skills and understanding of the culture of the organization are not less important than the traditionally emphasized technological job of turning data into insights.

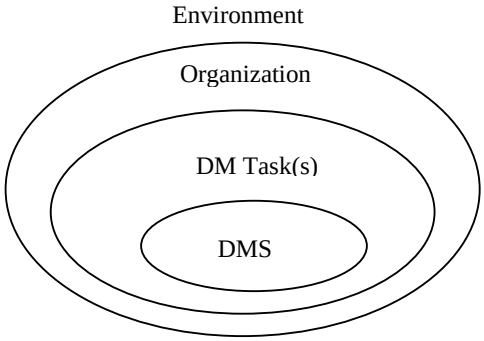

Fig. 8. DMS in the kernel of an organization

Hermiz [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) 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. Thus it seems possible that there are also some DMS specific questions that have not maybe been considered from those viewpoints in the IS discipline.

Lin in Wu *et al.* [41] 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 [16] 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.

4.2 The DMS artifact development environment

If a stated research problem includes a verb like introduce, improve, maintain, cease, extend, correct, adjust, enhance and so on, the study likely belongs to the area of constructive research. These are the kind of actions that researchers in the area of DM perform, when they are developing new theories and their applications as new artifacts to the use of persons and organizations. When a researcher him/herself is acting also as a change agent developing the artifact to an organization he is applying the action research approach.

But how to conceive, construct, and implement an artifact? It is obvious that in

order to construct a good artifact background knowledge is needed both about the artifact's components, that are the basic data mining techniques in the DM context and about components' cooperation, that are commonly selection and combination techniques in the DM context. Beside this the developer needs to have enough background knowledge about the human and organizational environment where the artifact is going to be applied. As discussed in Section 3 the design science approach is the one concentrating on this kind of research questions. Both these: the action research and design science approach to artifact creation and the evaluation process [25] are presented in Figure 9.

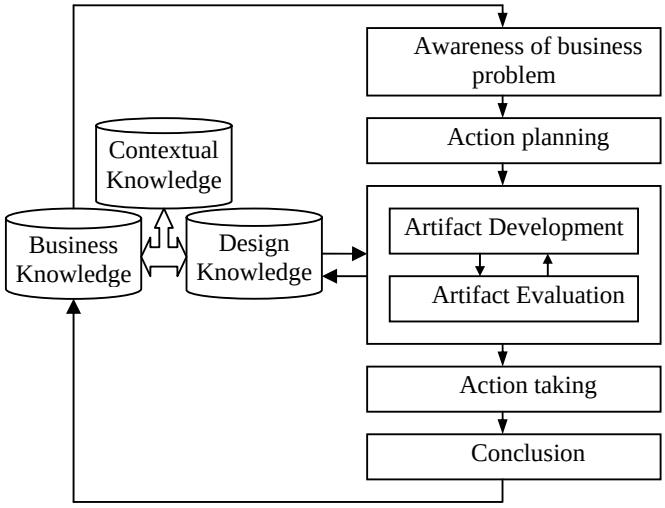

Fig. 9. The action research and design science approach to artifact creation

As discussed in Section 3 with Nunamaker's multimethodological approach it is essential that artifacts developed are also experimented with and analyzed using observation type of research. The evaluation process is a key part of any constructive research. This is also true when the artifact developed during research is a DMS (or its prototype). Usually, the experimental approach is used to evaluate a DM artifact. The experimental approach, however, can be beneficial for theory testing and can result in new pieces of knowledge thus contributing to the theory-creating process, too.

A 'goodness' criterion of a built theory or an artifact can be multidimensional and it is sometimes difficult to be defined because of mutual dependencies between the compromising variables. However, it is more or less easy to construct a criterion based on such estimates as accuracy of a built model and its performance. On the other hand, it is more difficult or even impossible to include into a criterion such important aspects as interpretability of the artifact's output because estimates of such kind are usually subjective and can be evaluated only by the end-users of a system. This does not eliminate the necessity to research also these topics which are important for users to see the results having relevance.

4.3 New DM research framework

Heavner et al. [20] presented a conceptual framework for understanding, conducting and evaluation of the IS research. We adapt their framework to the context of DM research (see Figure 10). The framework combines together the behavioral-science and design-science paradigms and shows how research rigor and research relevance can be explained, evaluated, and balanced.

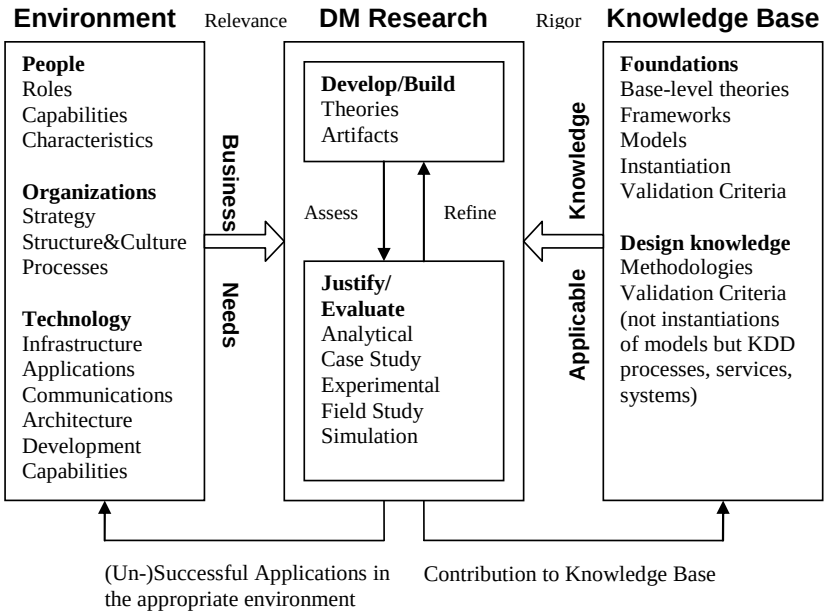

Fig. 10. New research framework for DM research (adapted from [20])

We follow Hevner at al. [20] with the description of the figure, emphasizing issues important in DM. The environment defines not only the data that represents the problem to be mined but people, (business) organizations, and their existing or desired technologies, infrastructures, and development capabilities. Those include the (business) goals, tasks, problems, and opportunities that define (business) needs, which are assessed and evaluated within the context of organizational strategies, structure, culture, and existing business processes. Those research activities that are aimed at addressing business needs contribute to the relevance of research.

Driven by the business needs, DM research can be conducted in two complementary phases. Behavioral science would guide research through the *development* and *justification* of theories that describe, explain or predict some phenomena associated with the business need being addressed. Design science enables the *building* and *evaluation* of artifacts being developed to address the business need. It is generally accepted that the goal of behavioral science research is truth and the goal of design science research is utility. However, Hevner at al.

[20] were likely the first who argued that truth and utility are inseparable – "truth informs design and utility informs theory" [20, p. 80]. Hevner at al. [20] conclude that "an artifact may have utility because of some as yet undiscovered truth. A theory may yet to be developed to the point where its truth can be incorporated into design. In both cases, research assessment via the justify/evaluate activities can result in the identification of weaknesses in the theory or artifact and the need to refine and reassess. The refinement and reassessment process is typically described in future research directions." [20, p. 80]

The knowledge base provides foundations and methodologies for research (and development) activities. Prior DM research and development and results from reference disciplines (statistics, machine learning, AI, etc.) provide foundational theories, frameworks, models, methods, techniques and their instantiations used in the develop/build phase of research. Methodologies should provide guidelines and techniques for the justify/evaluate phase. Rigor is achieved by appropriately applying existing foundations and methodologies.

5. Discussion and Conclusions

In this paper we first considered several existing frameworks for DM and their advantages and limitations. Second, we considered a traditional IS framework and two sub-frameworks: one for the IS user environment and another for the IS development environment. Based on these two we suggested our new research framework for DM. It imports research questions and topics from the IS discipline into the DM area trying to take benefit of the fact that the long developed IS discipline can help the maturing DM research area to raise the relevance of its research and thus its practical importance for people, organizations, and their surroundings.

Figure 11a presents our understanding of the current situation with DM research. The left triangle presents the current DM practice situation where almost merely the relevance aspects, i.e. utility are dominating. The right triangle presents the current DM research situation that is heavily dominated by rigor aspects and almost no attention is paid to DM research relevance. The lower arrow between DM research and practice is solid because some amount of rigor DM research results are flowing to the practice at least through software applications. The upper arrow is dashed because our understanding of the situation is that too seldom DM research takes practice related aspects into account and thus the exchange between DM practice and DM research is not as fruitful as it might be.

Even those relevance issues that are recognized within community of DM practitioners or let us say the (current or potential) users of DM systems, DM solutions and DM services, are not studied appropriately from scientific point of view and therefore we can rarely see the transfer of scientific knowledge (and in many cases even valuable feedback) from DM practice to DM research.

Thus, our believe is that within DM research community there should be DM research dealing purely with rigor issues, DM research dealing mostly with relevance issues and, likely the most challenging part of DM research efforts dealing with rigor/relevance aspects (Figure 11.b). With regard to this belief we recognize two important aspects: (1) those who practice DM should be well-motivated to share their expertise and scientific insights into relevance issues in

DM, and that is not less important, (2) DM research community should be interested in conducting and publishing *academic* research of relevance issues in DM. However, our analysis show that currently DM research often does take relevance into account only from empirical research point of view with regard to possible variety of dataset characteristics, but in most of the cases does not account for many important environment aspects (people, organization etc), i.e. relevance concept originating from design science.

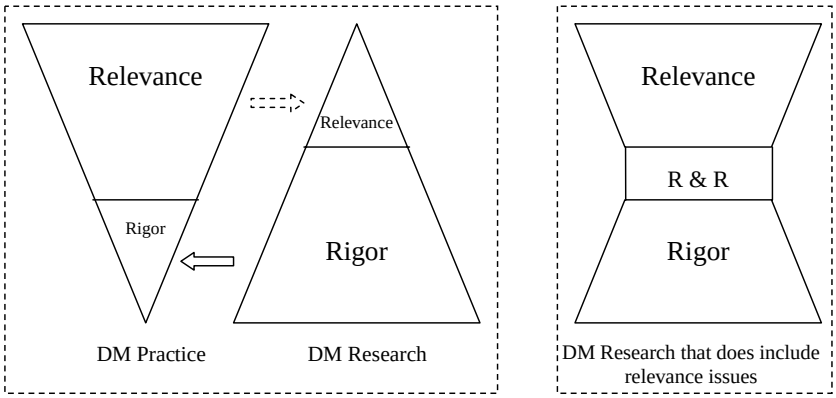

Fig. 11. Rigor and relevance aspects of DM research.

We considered DMSs as a special kind of ISs which have not yet been considered closely enough, in our opinion, from the use and development perspectives. After discussing these two DMS environments, we presented our new DM research framework, which aims at better balancing between the rigor and relevance constituents of research also in the DM area.

In this work we have not provided any examples to demonstrate the applicability of the proposed framework. We have not tried also to describe all the essential issues at the very detailed level, leaving this maturation for further research. However, we believe that our work could be helpful in turning the focus of DM research into a more balanced direction. We see this important from the point of view of raising DM first among those technologies which are able to produce competitive advantage and later to be developed to be one of 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, maybe 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 workshops 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. We hope also that meta-level research in DM, directed to the study of current situation and

trends and possibilities of further development of the field (as our study does) will be recognized as important and valuable type of 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. We are thankful to the reviewers of this chapter for their valuable comments and suggestions.

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. Bensusan H. "Is machine learning experimental philosophy of science?" In *Proc. of ECAI'2000 Workshop on Scientific Reasoning in AI and Philosophy of Science*, 2000, pp. 9-14.
3. Boulicaut J., Klemettinen M., and Mannila H. "Modeling KDD Processes within the Inductive Database Framework". In *Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery*, Springer-Verlag, London, UK, 1999, pp. 293-302.
4. Burrell G., Morgan G. "Sociological paradigms and organizational analysis", Heinemann, London, 1979.
5. Carvalho D.R. and Freitas A.A. Evaluating the correlation between objective rule interestingness measures and real human interest. *Proc. European Conf. on Principles and Practice of Knowledge Discovery in Databases* (PKDD-2005). LNAI 3721, pp. 453-461. Springer, 2005
6. Coppi R. "A theoretical framework for Data Mining: "Informational paradigm", *Computational Statistics and Data Analysis 38*, 2002, pp. 501-515.
7. Coppock D. S. "Data Mining and Modeling: So You have a Model, Now What?" *DM Review Magazine*, Feb 03.
8. CRISP-DM: 1.0 *Step-by-step DM guide*, SPSS Inc.
9. 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.
10. DeLone W., McLean E.R. "Information Systems Success: The Quest for the Dependent Variable", *Information Systems Research 3(1)*, 1992, pp. 60-95.
11. 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
12. 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
13. Fayyad U.M. "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert 11(5)*, 1996, pp. 20-25
14. Fayyad U.M., Uthurusamy R "Evolving data into mining solutions for insights". *Communications of the ACM 45(8)*, 2002, pp. 28-31
15. Friedman J. "Data Mining and Statistics: What's the connection?" In *D. Scott (Ed.) Proc. 29th Symposium on the Interface*, 1999.
16. Giraud-Carrier C. *Success Stories in Data/Text Mining*, Brigham Young University, 2004 (An updated version of an ELCA Informatique SA White Paper)
17. Hand D.J. "Data mining: statistics and more?" *The American Statistician, 52*, 1998, pp. 112-118.
18. Hand D.J. "Statistics and data mining: intersecting disciplines", *SIGKDD Explorations*

1, 1999, pp. 16-19.
19. Hermiz K.B., "Critical Success Factors for Data Mining Projects", *DM Review Magazine*, February 1999.
20. Hevner A., March S., Park J., Ram S. Design Science in Information Systems Research, MIS Quarterly 26(1), 2004, pp. 75-105.
21. 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
22. Imielinski T., Mannila H. "A database perspective on knowledge discovery", *Communications of the ACM 39(11)*, 1996, pp. 58-64.
23. Ives B., Hamilton S., Davis G. "A Framework for Research in Computer-based Management Information Systems", *Management Science 26(9)*, 1980, pp. 910-934.
24. Järvinen P. 2002. Research methods. Tampere. Opinpaja. (http://www.uta.fi/~pj/)
25. Järvinen P. 2005. Action Research as an approach in design science. TR D-2005-2. University of Tampere, Department of Computer Science.
26. Kleinberg J., Papadimitriou C., and Raghavan P. "A Microeconomic View of Data Mining," *Data Mining and Knowledge Discovery 2(4)*, 1998, pp. 311-324.
27. Lin T.Y. "Data Mining: Granular Computing Approach" In N.Zhong, L.Zhou (Eds.) *Proc. 3rd Pacific-Asia Conference, Methodologies for Knowledge Discovery and Data Mining*, LNCS 1547, 1999, pp. 24-33.
28. Lin T.Y., Granular Computing of Binary relations I: Data Mining and Neighborhood Systems. In: Polkowski and Skowron (Eds.) *Rough Sets and Knowledge Discovery*, Springer-Verlag, 1998, 107-121.
29. Lyytinen, K., 1987, Different perspectives on information systems: problems and solutions. ACM Computing Surveys, 19(1), 5-46.
30. Mannila H. "Theoretical Framework for Data Mining", *SIGKDD Explorations 1(2)*, 2000, pp. 30-32.
31. Mason R.O. Experimentation and knowledge – A paradigmatic perspective, Knowledge: Creation, Diffusion, Utilization 10(1), pp. 3-24.
32. Mehta M., Rissanen J., Agrawal R. "MDL-Based Decision Tree Pruning", In *Proc. KDD 1995*, 1995, pp. 216-221
33. Michalski R.S. "Seeking Knowledge in the Deluge of Facts", *Fundamenta Informaticae 30*, 1997, pp. 283-297.
34. Nunamaker W., Chen M., Purdin T. "Systems development in information systems research", *Journal of Management Information Systems 7(3)*, 1990-91, 89-106.
35. 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.
36. Pechenizkiy M., Puuronen S., Tsymbal A. 2005. Why Data Mining Does Not Contribute to Business? In: C.Soares et al. (Eds.), *Proc. of Data Mining for Business Workshop, DMBiz* (ECML/PKDD'05), Porto, Portugal, pp. 67-71.
37. 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 (Invited paper).
38. Reinartz, T. 1999, *Focusing Solutions for Data Mining.* LNAI 1623, Berlin Heidelberg.
39. Tan P., Kumar V., Srivastava J. Selecting the right objective measure for association analysis. *Information Systems 29(4)*, 2004, pp. 293-313.
40. Vapnik V. "The nature of statistical learning". Springer, NY, 1995.
41. Wu X., Yu P., Piatetsky-Shapiro G., et al. "Data Mining: How Research Meets Practical Development?" *Knowledge and Information Systems 5(2)*, 2000, pp. 248 – 261.

42. Yao J.T., Yao Y.Y. “A granular computing approach to machine learning”, *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'02)*, Singapore, November 18-22, 2002, pp732-736.
43. Zadeh L.A., Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, 90, 111-127, 1997.