On Conceptual Modeling of Data Mining

Yiyu Yao

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada, S4S 0A2
yyao@cs.uregina.ca

1. Introduction

The study of data mining has focused primarily on the mining algorithms and their applications, while relies its foundations on established fields, such as logic, cognitive science, statistical analysis, machine learning, databases, and so on. Motivated by the practical needs of specific types of real world data analysis problems, many mining algorithms are designed and studied. They include association rule mining, classification rule mining, exception and peculiarity rule mining, sequence mining, stream mining, text mining, web mining, and others. A review of data mining literature suggests that there does not exist a well-accepted and non-controversial conceptual framework. A lack of conceptual modeling may jeopardize further development of data mining.

It is perhaps the time to study data mining systematically as a branch of computer science. The chapter attempts to make a contribution to this trend. Specifically, we discuss a few foundational issues related to the conceptual modeling of data mining. We summarize our research results in the past few years [Yao01, Yao03, Yao04, YZ04, YZM03, YZZ04]. By putting them in a more coherent manner, we add new understanding and more insights into data mining.

Our discussions are unique and differ from existing studies in several perspectives. First, we treat data mining as a field of study and emphasize the study of the nature, the scope, and philosophical foundations of data mining. We stress on the understanding of data mining as a scientific inquiry, in addition to simply empirical investigations. We pay more attention to the effectiveness of data mining methods, rather than only to the efficiency. Second, we view data mining in a wide context of scientific research, in terms of their goals, processes and methods. Third, we search for a unified and general framework, or at least general principles and guidelines, rather than a family of isolated algorithms. The framework aims at finding answers to what and why questions, as well as how questions. Forth, with the help of conceptual modeling, we attempt to move beyond trial and error, or ad hoc, applications of data mining algorithms, which dominates most current applied studies of data mining.

Data mining is relatively a new field and has not yet formed its own theories, views and culture. A good starting point may be to examine the philosophy and principles proven to be successful in other established fields and branches of computer science, and to apply them to data mining. The explorations of this chapter are based on this underlying assumption. It draws extensively results from a number of fields. We divide the discussions into three parts. In the first part, we argue for the study of the conceptual modeling of data mining. The philosophical, conceptual understanding of data mining may shed new light on data mining research. It helps us to resolve the difficulties with existing data mining research. One simply cannot expect a continuous growth and development of a field without a solid foundation. The establishment of a foundation indicates the maturity of a field. In the second part, we present a comparative analysis of scientific research and data mining [YZ04]. By showing their connections, results from scientific research methods can be immediately applied to data mining. The comparative study provides us a new view of data mining, namely, data mining systems can be viewed as research support systems [Yao03a]. In the third part, we present a three-layered conceptual framework, which consists of the philosophy layer, the technique layer and the application layer [Yao03, YZZ04]. Each layer addresses different types of fundamental questions regarding data mining, and jointly they give a complete characterization of the field. By separating fundamental issues into different levels, the three-layered framework enables us to examine them more conveniently and systematically. It also helps us to observe problems in existing data mining studies, which are difficult to see otherwise.

The investigation of this chapter is exploratory in natural. The aim is to give a broad perspective of the problem at a higher level without bearing down by unnecessary details of any particular algorithms. We hope the discussion will stimulate some researchers to look further into the vital issues. The discussion offers some of the possible solutions, but not necessarily the best solutions. For example, the three-layered framework is not necessarily a most suitable conceptual model, or better than existing models. The framework is important in the sense that it offers an alternative view, which deserves its due attention. For a fully understanding, and further development, of data mining, one must investigate views complementary to the contemporary algorithm-dominated views and application-dominated views.

2. Conceptual Modeling

In justifying the needs for conceptual modeling and foundations of data mining, it is necessary first to present the current status of the field and to identify the associated difficulties. Potential solutions can then be sought.

2.1. A Brief Summary of Data Mining Research

The volume of research activities and its fast growth speed perhaps justify data mining as a solid research field on its own rights. A commonly used definition of data mining defines it as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data” [FPS96]. Even from this simple definition, we can observe a few perspectives. Each of them attempts to capture the intuitive notions of “pattern”, “process”, “validity”, “novelty”, and “understandability”. The definition concisely summarizes three views of data mining, the function-oriented, the procedure-oriented, and the application-oriented views. By adding the theory-oriented views [Man00, Yao01, Yao03], we have at least four dominant views of data mining [YZZ04].

The function-oriented views focus on the requirements and goals of data mining tasks. That is, data mining attempts to extract knowledge from data. Such goal-driven approaches establish a close link between data mining research and real world applications. Due to the diversity of data and different forms of knowledge, data mining inherently deals with difficult problems. One needs to consider different data mining systems with different functionalities and for different purposes, such as text mining, web mining, sequence mining, and temporal data mining. Two main objectives of data mining have been identified as prediction and description. Prediction involves the use of some variables to predict the values of some other variables, and description focuses on patterns that describe the data [FPS96].

The theory-oriented views concentrate on the theoretical studies of data mining, in relation to the other disciplines. Many theories and models of data mining have been proposed, critically investigated and examined [FPS96, Man00, Yao01, Yao03, YZM03]. Fields contributing to the theoretical study include logic, statistics, machine learning, databases, pattern recognition, visualization, and many others. There is also a need for the combination of existing theories. For example, some efforts have been made to bring logic, utility and measurement theory, concept lattice and knowledge structure, and other mathematical and logical models into the data mining models [Che02, LHOL03, LL02, LSPL04, LO02, Man00, XR02, Yao01, Yao03, YZZ04].

The procedure-oriented views cover two parts, namely, data mining algorithms and multiple phases of data mining process. Data mining algorithms deal with specific methods for mining particular types of knowledge. A multi-phase process describes the main steps involved in data mining. It normally consists of data selection, data preprocessing, data transformation, pattern discovery, pattern evaluation, and result explanation [FPS96, FPS96a, Man97, YZ04, YZM03, ZLO01]. In addition, the components of the process can be dynamically organized [ZLO01].

The application-oriented views deal with the utilization of data mining algorithms and techniques in various domains. Applications are in fact the driving market of data mining research. To a large extent, the research of data mining is motivated by practical needs of extracting useful knowledge from huge collected data in the first place.

While some progresses have been made with respect to the theory-oriented view, the mainstream research is concentrated on the other three views. The study of the foundations of data mining attempts to correct such an uneven development.

From the literature of data mining, one can also observe a few trends, representing four directions of growth of data mining research. One dimension is characterized by the size of databases. It becomes a common practice to apply a data mining algorithm to huge datasets of ever-increasing sizes. To overcome the difficulties associated with the sizes, many studies attempt to address the scalability of algorithms and speed-up of existing algorithms. This leads to an over-emphasis on the efficiency of mining algorithms. The types of data, in terms of format and content, define another dimension of data mining research. One moves from association mining to sequence mining, to stream mining, to text mining, and to web mining. The third dimension is along the application domains. Many studies apply existing methods into new domains, where data mining techniques had not been attempted. The forth dimension, which is perhaps more important than the other three dimensions, is defined by the types of knowledge. New data mining algorithms are introduced everyday, attempting to discover new types of knowledge. It should also be pointed out that many new data mining algorithms are only slight modifications and extensions of existing algorithms from other fields. For example, some text mining algorithms are in principle classical information retrieval algorithms.

A common feature of the four dimensions of growth is the expansion to a new territory, that is, a larger dataset, a new type of data, a new domain, or a new type of knowledge. Obviously, such a growth increases the volume of research. A crucial question is if the increase in quantities leads to a deeper understanding of the field. The answer to this question may not be a really yes. It is true that we understand data mining better than a decade ago. With more than a decade development, we have more algorithms and more applications. They, unfortunately, do not necessarily increase our understanding of the problem itself, that is, transforming data into information, information into knowledge, and knowledge into wisdom. The conceptual modeling of data mining may offer some help in achieving this goal.

2.2. Motivations for Conceptual Modeling

In order to see the needs for the study of foundations of data mining, let us first quote the following comments from Salthe [Sal85]:

> “Functioning as a scientist means functioning within the rules of a game learned during an apprenticeship in which examination of the philosophic foundations of the game plays a characteristically tiny role. One strives to become a member, not to potentially undermine the club by examining its structure from outside. Only when commitment to a way of life is secure is it possible for some to examine its foundations with sympathy. The result is that the typical young scientist is trained to measure, assuming that what he measures exists, and he is little cognizant of how little his measurements justify that working assumption. Justification, in fact, is not required as long as the science is flourishing, contributing its share to the social context. But, when it falters, we fall upon times of foundational reexamination, as with evolutionary biology today.”

Although the comments are made from an ecologist’s point of view, they are equally applicable to data mining research. They may explain why researchers do not examine the foundational issues, especially when the initial success of data mining is well pronounced and reported. A more important point is that we perhaps should examine foundational issues early, rather than waiting for the time when a lack of foundations restricts the growth of the field.

Foundational investigations enable us to gain a conceptual understanding of a field. As pointed out by Simpson [Sim96], "The foundations of X are not necessarily the most interesting part of field X. But foundations help us to focus on the conceptual unity of the field, and provide the links which are essential for applications and for integration into the context of the rest of human knowledge." Without a unified conceptual understanding, we can only have fragmented and local views of a field. For example, in the context of ecosystem, Salthe [Sal85] states, "The question typically is not what is an ecosystem, but how do we measure certain relationships between populations, how do some variables correlate with other variables, and how can we use this knowledge to extend our domain." The similar observations can be made for data mining research. More specifically, one is more interested in the algorithms for finding knowledge, but not what is knowledge and what is the knowledge structure. One is often more interested in a more implementation-oriented view or a concrete framework of data mining, rather than a conceptual framework for the understanding of the nature of data mining [YZZ04].

The discussion converges to an important conclusion. The requirement for conceptual modeling of data mining is no longer a luxury, but a necessity for further, healthy, sustainable development of the field. With proper conceptual modeling, one can gain more insights into knowledge extraction from data, instead of yet another mining algorithm or another application.

2.3. Foundations of Data Mining

There is an emerging interest in the foundations of data mining, which unfortunately did not receive enough attention until recently [Che02, Lin02, Man00, ML02, WZZH03, XR02, Yao01, Yao03, YZZ04], notably by a series of workshops initiated by Lin and colleagues [LHOL03, LL02, LSPL04, LO02]. The study of foundations of data mining deals with conceptual modeling of data mining as a field of scientific inquiry. It examines into the nature of data mining and the scope of data mining methods. It treats data mining as an integrated whole and a subject of study, rather than an isolated family of algorithms and applications. It studies the conceptual structures of data mining, which link its various notions [Yao01, Yao03].

In stating foundations of mathematics, Simpson [Sim96] makes explicit a few important points. First, human knowledge is conceptual, contextual, and hierarchical, which forms an integrated whole. Human knowledge is organized hierarchically into a tower or a partial ordering. The most fundamental concepts form the base or minimal elements of the ordering. Higher-level concepts are derived or defined based on lower-level concepts [Pei91]. Second, a field of study normally covers a part of the integrated whole of human knowledge. It is distinguished by a certain conceptual unity in the sense that the concepts of the field are closely related to each other and are sufficiently self-contained. Consequently, a field can be studied in isolation for some purposes. The conceptual unity usually is implied by the existence of a specific subject matter, i.e., the real-world object of study. Third, foundations of a field normally refer to a more-or-less systematic analysis of the most basic or fundamental concepts of the field. The framework of Simpson can be immediately applied to establish foundations of data mining [Yao03].

Many researchers also support conceptual modeling, based on knowledge structures, as a way to understand a field and to apply the results from the field. In the context of solving physics problems, Reif and Heller [RH82] state, “effective problem solving in a realistic domain depends crucially on the content and structure of the knowledge about the particular domain”. Knowledge about physics in fact specifies concepts and relations between them at various levels of abstraction. Furthermore, the knowledge is organized hierarchically, with explicit guidelines specifying when and how this knowledge is to be applied. Posner [Pos89] suggests that, according to the cognitive science approach, to learn a new field is to build appropriate cognitive structures and to learn to perform computations that will transform what is known into what is not yet known.

It is evident that the foundations of data mining can be established by focusing on a set of closely related concepts. The conceptual study makes explicit the conceptual knowledge structures of data mining. The hierarchical organization of data mining concepts provides an easy way to understand the description of data mining. In addition, guidelines, specifying when and how the knowledge of data mining can be used, must be studied. A systematic study of the basic notions and the knowledge structures of data mining will bring it into a field of study on its own rights.

2.3. Implications

The conceptual study focuses on a different level of understanding of data mining. It may lead to a powerful point of view, but may not immediately lead to a new algorithm or offer an improved algorithm. Its relevance to the applications of data mining may seem to be even more remote. Consequently, not enough attention is paid to conceptual studies. A lack of conceptual study may account for much of the misunderstanding and confusing of many fundamental issues, repeated research efforts, misuses of data mining algorithms, and fruitless pursue of certain types of research.

It should be realized that a powerful way of thinking, derived from conceptual studies, enables us to have an in-depth understanding of the field. This in turn leads to a proper conceptualization, formulation, and representation of the problems, and successful applications of the theories and techniques. We can avoid many pitfalls and be immune to many potential mistakes. It is exactly for such reasons that we pay attention to less studied conceptual modeling of data mining.

3. Data Mining and Scientific Research

Extracting knowledge from data or making sense out of data has been, and is still, a basic endeavor of any scientist. The term data is used here in a very broad sense, covering any format and any content. Categorically speaking, the tasks and methods explored in data mining are not out of the scope of scientific research. It is therefore constructive to examine data mining in a wide context of scientific research [YZ04].

3.1. Common Purposes and Goals

Scientific research is affected crucially by the perceptions and the purposes of science. Martella *et al.* [MNM99] summarize the main purposes of science, namely, to describe and predict, to improve or manipulate the world, and to explain our world around us. The results of the scientific research process provide a description of an event or a phenomenon. The knowledge obtained from research helps us to make predictions about what will happen in the future. Research findings are useful for us to make an improvement in the subject matter. Research findings can be used to determine the best or the most effective interventions to bring about desirable changes. Finally, scientists develop models and theories that account for a natural phenomenon.

An important question in scientific research is the perception of objectivity and subjectivity. It is true that the subject matter, the real-world object of study, may be an objective reality. However, our understanding depends largely on our choice of the description and the context in which we formulate the design and results of our observations, which is rather subjective [Pat73]. Both objectivity and subjectivity are mixed together to characterize scientific research.

Researchers in data mining have discussed goals similar to those of scientific research. For example, Fayyad *et al.* [FPS96a] present two high-level goals of data mining as prediction and description. In other words, it re-expresses some of the goals of science, in the context of data mining. Ling *et al*. [LCYC02] study the issue of finding optimal actions to increase profit. It in fact deals with the manipulation based on the discovered knowledge.

Based on the connections between scientific research and data mining, Yao *et al.* [YZ04, YZM03] consider the goal of explanation. A framework of explanation-oriented data mining is proposed, which focuses on constructing models accounting for data mining results. A unique feature of explanation-oriented mining is the re-collection and construction of explanation data. The original data may only show the occurrence of phenomenon. In order to construct models of explanation, it may be necessary to re-collect data that summarize other aspects of the problem. The re-collection of data and information is a typical method in scientific research. It is through the re-recollection of data that we make data mining to be an incremental process of scientific exploration.

3.2. Common Processes

Research is a highly complex and subtle human activity, which may be difficult, if not impossible, to formulate formally. Nevertheless, some lessons and general principles can be learnt from the experience of scientists. There are some basic principles and techniques that are commonly used in most types of scientific investigations. Granziano and Raulin [GR00] make a clear separation of research process and content:

> "The particular observations made vary from one discipline to another because each discipline is interested in observing and understanding different phenomena. But the basic processes and the systematic way of studying problems are common elements of science, regardless of each discipline's particular subject matter. It is the process and not the content that distinguishes science from other ways of knowing, and it is the content – the particular phenomena and fact of interest – that distinguishes one scientific discipline from another."

It is this common process that makes the investigation of research methods possible. We adopt the model of the research process of Garziano and Raulin [GR00], and combine it with other models [MNM99]. The basic phases and their objectives are summarized as follows:

- Idea-generation phase: to identify a topic of interest.
- Problem-definition phase: to precisely and clearly define and formulate vague and general ideas generated in the previous phase, and to identify a particular problem of study.
- Procedure-design/planning phase: to make a workable research plan by considering all issues involved.
- Observation/experimentation phase: to observe real world phenomenon, collect data, and carry out experiments.
- Data-analysis phase: to make sense out of the data collected.
- Results-interpretation/explanation phase: to build rational models and theories that explain the results from the data-analysis phase.
- Communication phase: to present the research results to the research community.

It is possible to combine several phases into one, or to divide one phase into more detailed steps. The division between phases is not a clear cut. The research process does not follow a rigid sequencing of the phases. Iteration of different phrases may be necessary [GR00].

The commonly used data mining process is similar, in nature, to the research process. The following is a summary of a typical data mining process [FPS96, FPS96a, YZ04, YZM03, ZLO01]:

- Data pre-processing phase: to select and clean working data.
- Data transformation phase: to change the working data into the required form.
- Pattern discovery and evaluation phase: to apply algorithms to identify knowledge embedded in data, and to evaluate the discovered knowledge.
- Explanation construction and evaluation phase: to construct plausible explanations for discovered knowledge, and to evaluate different explanations.
- Pattern presentation: to present the extracted knowledge and explanations.

There are two levels of connections between scientific research and data mining. If data mining is treated as a field of study, there is a parallel correspondence between the processes of scientific research and data mining. By substituting the subject of data mining, i.e., data and the associated concepts, into the general research process, one immediately obtains the correspondence. On the other hand, if data mining is treated literally, it deals with the data analysis phase of the research process. The data mining process is simply a lower level, more detailed process of data analysis.

3.3. Implications

The examination of data mining in a wider context of scientific research lends itself to a bi-directional interaction of data mining and scientific research, whose potential implications cannot be over-emphasized.

From the viewpoint of scientific research, it is possible to have an in-depth understanding of data mining. The four dimensions of growth of data mining research is indeed closely related to some aspects in the evolution of science in general, namely, the application of old ideas to new data or new types of data, the exploration of new scientific ideas, and the emergence of new branches of science. The experiences and results from the studies of research methods can be applied to data mining. Many different ways, methods, tools, and ideas have been experimented by scientists again and again in many different domains. This opens new doors to data mining research. In carrying out data mining tasks, one can borrow or adopt scientific research methods and ideas that have been used either explicitly or implicitly by scientists.

On the other hand, data mining systems can be used to support scientific research. In other words, data mining is considered to be part of research support systems [Yao03a]. In fact, data mining functionality has been incorporated into many intelligent systems such as knowledge management systems and decision support systems [TA01].

The examination implicitly suggests the power of a seamless integration of human and computer. The combination of scientific research methods and data mining will achieve the combined power of human and computer, i.e., the power of computer in fast data processing and analysis, and the power of human insights, intuitions, creativity, and analytical skills [Bev57]. The creation and growth of the web has amply demonstrated the implications of such a power [Ber99]. The success of data mining will again explore this power.

4. Multi-Level Modeling of Data Mining

The needs for a conceptual modeling of data mining, as well as its benefits, have been argued early. What left in this section is to actually construct such a conceptual framework. We first introduce Marr’s ideas of multi-level understanding of information processing systems [Mar82], and then discuss three levels for the understanding of data mining [Yao03, YZZ04].

4.1. Multi-level Understanding of Information Processing Systems

In his work on vision, Marr [Mar82] convincingly argues that a full understanding of an information processing system involves explanations at various levels or layers. A special feature of this hierarchical analysis is the consideration of computational issues, such as representation and process, and implementation. As a basic component of information processing systems, a process can be understood at three levels [Mar82].

The most abstract level deals with what the process does and why. One builds a theory that explains internal working principles of the process, and defines the operations by specifying constraints that must be satisfied by the process.

The second level deals with the realization of the process in an abstract way. One needs to choose a representation for the input and for the expected output of the process, and to specify an algorithm for the transformation from input to output. The choices of representation and algorithm are closely tied together. There usually exist many alternative representations. For a given representation, there are also many possible algorithms. A representation and an algorithm should be chosen so that advantages of the representation are fully exploited by the algorithm and, at the same time, the disadvantages of the representation are avoided.

The third level deals with the physical realization of the process. The devices that physically realize a process may not be unique. The advances in technologies imply that the same process may be implemented again with the invention of new physical devices.

The three-level understanding of a process can be generalized to an information processing system. While the first level answers questions of what and why, the other two levels answer two types of questions of how. More specifically, adopted from Figure 1-4 in Marr's book, they are given by [Mar82]:

- Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
- Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input an output, and what is the algorithm for the transformation?
- Hardware implementation: How can the representation and algorithm be realized physically?

Investigation at the computational theory level is independent of representations, and investigations at representation and algorithm level is independent of physical devices. The levels are ordered and interpreted as levels of abstraction. There are in fact two implementation levels. The representation and algorithm level may be viewed as the logical implementation level, and the hardware implementation level as the physical implementation level.

4.2. A Three-layered Framework of Data Mining

A three-layered conceptual framework consists of the philosophy layer, the technique layer, and the application layer. The layered framework represents the understanding, discovery, and utilization of knowledge [Yao03, YZZ04]. The framework basically adds an application level to Sowa's [Sow84] information processing model, and at the same time taking consideration of Marr's [Mar82] multi-level understanding of information processing systems.

The philosophy layer: The philosophy layer investigates the basic issues of knowledge. One attempts to answer the fundamental question, namely, what is knowledge? There are many related issues, such as the representation of knowledge, the expression and communication of knowledge in languages, the storage and processing of knowledge in mind, the relationship between knowledge in abstract, in the mind and in the external real world, and the classification and organization of knowledge [Sow84]. Philosophical study of data mining serves as a precursor to technology and applications. It generates knowledge and the understanding of our world, with or without establishing operational boundaries of knowledge.

The technique layer: The technique layer is the study of knowledge discovery methods and their implementations in machine. Two levels can be further formed, i.e., the logical implementation and physical implementation. One attempts to answer the fundamental question, how to discover knowledge? Logical analysis and mathematical modeling are more relevant at the logical implementation level. One is interested in searching for new algorithms and improving existing algorithms. The physical implementation expresses knowledge discovery methods through programming languages, which involves the coding, storage and retrieval of information. The main streams of research in machine learning, data mining, and knowledge discovery have concentrated on the technique layer.

The application layer: The ultimate goal of knowledge discovery is to effectively use the discovered knowledge. At this layer, one needs to answer the question, how to utilize the discovered knowledge? This layer attempts to make explicit and precise the intuitive notions of usefulness and meaningfulness of discovered knowledge, based on domain specific background knowledge.

The three layers are relatively independent and loosely connected. The inner or lower layers establish a foundation for the outer or upper layers, while the outer layers may raise questions for the inner layers. This explicit division, although may be artificial, enables us to see the basic issues of data mining more clearly. Useful and insightful remarks can be made regarding the three-level framework. The results from philosophy level will provide guideline and set the stage for the algorithm and application levels. Philosophical study does not depend on the availability of specific techniques. More specifically, the existence of a particular type of knowledge, as well as its usefulness, is not determined by the existence of a mining algorithm for such knowledge. Knowledge is a matter of existence. Data mining algorithms merely reveal the knowledge embedded in the data. The technique level study is not constrained by a particular application. The existence of an algorithm does not necessarily imply that the discovered knowledge is meaningful and useful.

4.3. Implications

It may be debatable regarding the appropriateness of the three-level framework, such as the number of levels, the division between levels, the interaction of different levels, and the issues at each level. The usefulness of the framework is evident. It represents a three-level description of data mining characterized by three fundamental issues, namely, the understanding, discovery, and utilization of knowledge. Each of them is indispensable, and jointly they present a framework within which a multi-disciplinary study of data mining is possible.

A significant implication of the framework lies on its division of the understanding of a complex problem into different levels, which leads to a division of basic issues of data mining into levels. It also gives the proper context in which a particular type of questions can be answered. One can clearly see the scope and limitations of various types of data mining research.

As an example of illustration, let us consider the notion of usefulness. It seems impossible to answer satisfactorily the question about the usefulness of a type of knowledge at the philosophy level or technique level. This question can only be answered at the application level, thanks to the domain specific background knowledge. The general discussion of the usefulness of knowledge, in terms of its novelty, validity or our awareness, in terms of its properties such as the complexity or structures, or in terms of the complexity of the discovery process, is perhaps ill conceived from the very beginning. A piece of commonly known knowledge may find new applications, and is therefore practically useful. A simple or trivial process may also turn out useful knowledge from data. The usefulness of knowledge is usually conditioned by one's background knowledge and the time of application.

In medical science, the effectiveness and usefulness of a new medicine is normally tested extensively on real world objects. This typically consists of two controlled groups, and the use of the medicine in one group and not in the other. Only after a long time of observations and comparisons of the two groups, one can make tentative conclusion of the effectiveness of the medicine. Similar types of tests are also conducted in the field of education, where the effectiveness of a teaching method is evaluated. It immediately follows that the similar testing method should have been used in data mining research, especially in business-oriented applications. For example, one can measure the usefulness of a piece of discovered knowledge in terms of profits it brings to an organization. Unfortunately, we do not see reports of such kind. Instead, we see a huge volume of tests based on some artificially defined criteria whose relevance to the real world problem is not clearly demonstrated.

It may be argued that the similar testing method can be used to evaluate the usefulness of the conceptual modeling of data mining, or in particular the three-level framework. In fact, initial classroom instruction does show that students grasp fast the basic issues of data mining with the three-level framework. Moreover, students can better formulate a research problem, in terms of its scope, basic issues, and potential solutions, with the help of multi-level conceptual description.

5. Concluding Remarks

In this chapter, we elaborate on several issues in the conceptual modeling of data mining. Since detailed description of data mining, in terms of theory, techniques, and algorithms, can be found easily in the literature, we only touch it briefly and superficially. Instead, we concentrate on a high-level conceptual modeling of data mining. An exposition of data mining is made in a wide context of scientific research. A three-layered conceptual framework is used for the understanding of data mining as a field of study. Our aim is to stimulate more researchers to look into conceptual modeling, a less studied, but extremely crucial, area of data mining research.

The content of the chapter is both old and new. It is old in the sense that many results are drawn from other fields, including cognitive science, philosophy of science, research methods, education, hierarchy theory, granular computing, and information processing systems. It is also new in the sense that these results are associated with new, domain specific meaning and are applied to data mining. In particular, we review the relevant results, interpret them, and apply them to data mining.

In some sense, this chapter is a product of a human text mining process, with the support of information retrieval systems. Some existing text mining systems are of little help in the preparation of this chapter. This mainly stems from either their special-purpose nature or a lack of support for human interaction in order to integrate knowledge in different domains. Furthermore, the materials of this chapter are results of in-depth logical reasoning, which apparently is not supported by current generation text mining systems.

Finally, the view presented in this chapter is complementary to the algorithm-dominated and application-dominated views of data mining. This view is perhaps not as important as the philosophy it promotes, data mining as a field of study can be better understood from multiple, diversified, and different views.

Acknowledgements

This chapter is based on a talk given at a Workshop on Machine Learning and Applications held at Fudan University (November 5-7, 2004). The author thanks Professors Jue Wang, Zhi-Hua Zhou, Aoying Zhou, and Ruqian Lu for the kind invitation. The author thanks Professor Ning Zhong, Yan Zhao and Yaohua Chen for their continued collaboration and constructive comments on data mining. The author thanks NSERC Canada for the partial support of this research.

Reference

[Che02] Chen, Z. The three dimensions of data mining foundation, Proceedings of IEEE ICDM’02 Workshop on Foundations of Data Mining and Knowledge Discovery, 119-124, 2002.

[Ber99] Berners-Lee, T. (with Fischetti, M.) Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web by Its Inventor, HarperSanFrancisco, 1999.

[Bev57] Beveridge, W.I.B. The Art of Scientific Investigation, Vintage Books, New York, 1957.

[FPS96] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. Data mining to knowledge discovery: an overview, in: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.), AAAI/MIT Press, 1-34, 1996.

[FPS96a] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. From data mining to knowledge discovery in databases, AI Magazine, 17, 37-54, 1996.

[GR00] Graziano, A.M. and Raulin, M.L. Research Methods: A Process of Inquiry, 4th edition, Allyn and Bacon, Boston, 2000.

[LCYC02] Ling, C.X., Chen, T., Yang, Q. and Cheng, J. Mining optimal actions for profitable CRM, Proceedings of IEEE International Conference on Data Mining, ICDM 2002, 767-770, 2002.

[LHOL03] Lin, T.Y., Hu, X.H., Ohsuga, S. and Liau, C.J. (eds.) Proceedings of IEEE ICDM'03 Workshop on Foundation of New Directions in Data Mining, 2003.

[Lin02] Lin, T.Y. Mathematical foundation of association rules – mining associations by solving integral linear inequalities, Proceedings of IEEE ICDM'02 Workshop on Foundation of Data Mining and Knowledge Discovery, 81-88, 2002.

[LL02] Lin, T.Y. and Liau, C.J. (eds.) Proceedings of the PAKDD'02 Workshop on Foundations of Data Mining, Communications of Institute of Information and Computing Machinery, 5 (2), 2002.

[LO02] Lin, T.Y. and Ohsuga, S. (eds.) Proceedings of IEEE ICDM'02 Workshop on Foundation of Data Mining and Knowledge Discovery, 2002.

[LSPL04] Lin, T.Y., Smale, S., Poggio, T, and Liau, C.J. (eds.) Proceedings of IEEE ICDM'04 Workshop on Foundations of Data Mining, 2004.

[Man97] Mannila, H. Methods and problems in data mining, Proceedings of International Conference on Database Theory, 41-55, 1997.

[Man00] Mannila, H. Theoretical frameworks for data mining, SIGKDD Explorations, 1, 30-32, 2000.

[Mar82] Marr, D. Vision, A Computational Investigation into Human Representation and Processing of Visual Information, W.H. Freeman and Company, San Francisco, 1982.

[ML02] Matloff, N. and Lin, T.Y. Toward statistical foundation for data mining, Proceedings of IEEE ICDM'02 Workshop on Foundation of Data Mining and Knowledge Discovery, 125-130, 2002.

[MNM99] Martella, R.C., Nelson, R. and Marchand-Martella, N.E. Research Methods: Learning to Become a Critical Research Consumer, Allyn and Bacon, Boston, 1999.

[Pat73] Pattee, H.H. Unsolved problems and potential applications of hierarchy theory, in: Hierarchy Theory, The Challenge of Complex Systems, Pattee, H.H. (ed.), George Braziller, New York, 129-156, 1973.

[Pei91] Peikoff, L. Objectivism: the Philosophy of Ayn Rand, Dutton, New York, 1991.

[Pos89] Posner, M.I. (ed.), Foundations of Cognitive Science, Preface: learning cognitive science, The MIT Press, Cambridge, Massachusetts, 1989.

[RH82] Reif, F. and Heller, J. Knowledge structure and problem solving in physics, Educational Psychologist, 17, 102-127, 1982.

[Sal85] Salthe, S.N. Evolving Hierarchical Systems, Their Structure and Representation, Columbia University Press, 1985.

[Sim96] Simpson, S.G., What is foundations of mathematics? 1996, http://www.math.psu.edu/simpson/hierarchy.html (accessed November 21, 2003).

[Sow84] Sowa, J.F., Conceptual Structures, Information Processing in Mind and Machine, Addison-Wesley, Reading, Massachusetts, 1984.

[TA01] Turban, E. and Aronson, J.E. Decision Support Systems and Intelligent System, Prentice Hall, New Jersey, 2001.

[WZZH03] Wang, J., Zhao, M., Zhao, K. and Han S. Multilevel data summarization from information system: a "rule + exception" approach, AI Communications, 16, 17-39, 2003.

[XR02] Xie, Y. and Raghavan, V.V. Probabilistic logic-based characterization of knowledge discovery in databases, Proceedings of IEEE ICDM'02 Workshop on Foundation of Data Mining and Knowledge Discovery, 107-112, 2002.

[Yao01] Yao, Y.Y. Modeling data mining with granular computing, Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC 2001), 638-643, 2001.

[Yao03] Yao, Y.Y. A step towards the foundations of data mining, in: Data Mining and Knowledge Discovery: Theory, Tools, and Technology V, Dasarathy, B.V. (ed.), The International Society for Optical Engineering, 254-263, 2003.

[Yao03a] Yao, Y.Y. A framework for web-based research support systems, Proceedings of 27th International Computer Software and Applications Conference (COMPSAC 2003), 601-606, 2003.

[Yao04] Yao, Y.Y. Concept formation and learning: a cognitive informatics perspective, Proceedings of International Conference on Cognitive Informatics, 42-51, 2004.

[YZ04] Yao, Y.Y. Explanation-oriented data mining, manuscript, 2004.

[YZM03] Yao, Y.Y., Zhao, Y. and Maguire, R.B. Explanation-oriented association mining using a combination of unsupervised and supervised learning algorithms, Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence, LNAI 2671, 527-531, 2003.

[YZZ04] Yao, Y.Y., Zhong, N. and Zhao, Y. A three-layered conceptual framework of data mining, Proceedings of IEEE ICDM’04 Workshop on Foundations of Data Mining, 205-212, 2004.

[ZLO01] Zhong, N., Liu, C. and Ohsuga, S. Dynamically organizing KDD processes, International Journal of Pattern Recognition and Artificial Intelligence, 15, 451-473, 2001.