Data Mining: How Research Meets Practical Development?

Xindong Wu[1], Philip S. Yu[2,1], Gregory Piatetsky-Shapiro[3], Nick Cercone[4], T.Y. Lin[5], Ramamohanarao Kotagiri[6], and Benjamin W. Wah[7]

[1] Department of Computer Science, University of Vermont
Burlington, VT 05405, USA

[2] IBM T.J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA

[3] KDnuggets
Brookline, MA, USA

[4] School of Computer Science, University of Waterloo
200 University Avenue, Waterloo, Ontario, Canada N2L 3G1

[5] Department of Mathematics and Computer Science
San Jose State University
San Jose, CA 95192, USA

[6] Department of Computer Science and Software Engineering
The University of Melbourne
Parkville, VIC 3010, Australia

[7] Computer Systems Research Laboratory
University of Illinois, Urbana-Champaign
1308 West Main Street, Urbana, IL 61801, USA

Abstract

At the 2001 IEEE International Conference on Data Mining in San Jose, California on November 29 - December 2, 2001, there was a panel discussion on how data mining research meets practical development. One of the motivations for organizing the panel discussion was to provide useful advice for industrial people to explore their directions in data mining development. Based on the panel discussion, this paper presents the views and arguments from the panel members, the Conference Chair and the Program Committee Co-Chairs. These people as a group have both academic and industrial experiences in different data mining related areas such as databases, machine learning, and neural networks. We will answer questions such as (1) how far is data mining from practical development, (2) how data mining research differs from practical development, and (3) what are the most promising areas in data mining for practical development?

1 Introduction

The IEEE International Conference on Data Mining (ICDM) provides a leading international forum for the sharing of original research results and practical development experiences among researchers

and application developers from different data mining related areas such as machine learning, automated scientific discovery, statistics, pattern recognition, knowledge acquisition, soft computing, databases and data warehousing, data visualization, and knowledge-based systems. The conference seeks solutions to challenging problems facing the development of data mining systems, and shapes future directions of research by promoting high quality, novel and daring research findings. In addition to business oriented data mining, ICDM has an equal emphasis on engineering, scientific, and medical data for which domain knowledge plays a significant role in knowledge discovery and refinement.

Since ICDM 2001 was held in San Jose, the heart of the Silicon Valley, the conference organized a panel discussion on how research meets practical development, aiming to provide useful advice for development people in Silicon Valley in particular and the data mining industry in general on practical directions of data mining. Below is some background information about each of the co-authors who have contributed towards the panel discussion and this paper.

Nick Cercone. Nick Cercone is Professor and Past Chair of Computer Science at the University of Waterloo (1997-2001). From 1993 until 1997 he was Associate Vice President (Research), Dean of Graduate Studies and International Liaison Officer at the University or Regina. Formerly he was Director of the Centre for Systems Science at Simon Fraser University (1987-1992) and Chairman of the School of Computing Science (1980-1985) at Simon Fraser.

Cercone's research interests include natural language processing, knowledge-based systems, knowledge-discovery in databases, data mining, and design and human interfaces.

Cercone received the BS degree in Engineering Science from the University of Steubenville in 1968, the MS degree in Computer and Information Science from Ohio State University in 1970, and a Ph.D. degree in Computing Science from the University of Alberta in 1975. Cercone worked for IBM Corporation in 1969 and 1971 on design automation. Cercone has authored 200 refereed technical papers and books.

Ramamohanarao Kotagiri. Ramamohanarao (Rao) Kotagiri is the Department Head of Computer Science and Software Engineering at The University of Melbourne. He is a member of the Steering Committee of the ICDM. He has served on many international conferences including VLDB, ICDE, SIGMOD, DOOD, PAKDD as a program committee member. He has also served as a Program Chair for several international conferences including VLDB, PAKDD, and DOOD. His interests include machine learning, deductive database systems, and information retrieval. His current interests are mining interesting emerging patterns efficiently.

T.Y. Lin. Tsau Young (T. Y.) Lin received his Ph.D. from Yale University, and is now a Professor at San Jose State University. He is the Founding President of the International Rough Set Society, and the Founding Chair of the Special Interest Group on Granular Computing. He has served as a program committee member (including being a (co-)chair) for many conferences, special sessions and workshops. He has also served as an editor-in-chief (now retired), associate editor, member of editorial/advisory/review boards for several international journals.

His primary interests are in three areas: database and knowledge-base systems (approximate retrieval and reasoning, data mining, e-intelligence), database security (data mining and inference problems, Internet privacy and security), and novel computing methodologies (including fuzzy, granular, Petri net, and rough computing).

His current focus is on the foundations of data mining; he has published papers on fast algorithms for association rules using granular computing, attribute transformation, feature completion, semantics oriented data mining, and data mining on derived attributes.

Gregory Piatetsky-Shapiro. Gregory Piatetsky-Shapiro, Ph.D. is the President and Founder of KDnuggets, which provides consulting and recruiting services in the areas of data mining, bioin-

formatics, and business analytics for CRM (customer-relationship management). He is also the Editor of KDnuggets News, the leading online newsletter for the data mining and knowledge discovery community.

He is the founder of the Knowledge Discovery and Data Mining (KDD) conference series, having organized and chaired the first three KDD workshops and currently serves as a Director of ACM SIGKDD, the professional association of Data Miners. In 2000 he received the first SIGKDD Service Award for his contributions to the Data Mining and Knowledge Discovery community.

Benjamin Wah. Benjamin W. Wah has worked on various aspects of machine learning, including genetic algorithms and artificial neural networks, with applications in data mining, computer load balancing, VLSI circuit design and testing, financial engineering, and algorithm design. His current work is focused on mining and predicting stationary and noisy time series, using formal theory and methods in nonlinear discrete constrained optimization.

Wah was one of the three founders of the *IEEE Transactions on Knowledge and Data Engineering* in 1987 and served as its Editor-in-Chief between 1993 and 1996. He has chaired a number of international conferences in the past, several of them in the areas of intelligent systems and data engineering. He is the 2001 President of the IEEE Computer Society.

Xindong Wu. Xindong Wu is Professor and Chair of the Department of Computer Science at the University of Vermont. He is the founder of the IEEE International Conference on Data Mining and also *Knowledge and Information Systems (An International Journal)*. He is the Chair of the ICDM Steering Committee and the Conference Chair of ICDM 2001.

Wu served on the 1995, 1996, and 2002 Program Committees of the KDD conference (now ACM SIGKDD), and is on the Editorial Board of *Data Mining and Knowledge Discovery* since the journal's inception in 1997. He is the Founding Chair (April 1998 - April 2001) of the Steering Committee for the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). He has published extensively in data mining and knowledge-based systems.

Philip Yu. Philip S. Yu is with the IBM Thomas J. Watson Research Center and currently Manager of the Software Tools and Techniques group, which focuses on data mining and optimization algorithms. He has published extensively on various data mining techniques, including clustering, classifications, association rule mining, and sequential pattern discovery. He has also contributed to various data mining application areas, including personalization, Web mining, knowledge management and bioinformatics. Dr. Yu has published more than 320 papers in refereed journals and conferences. He holds or has applied for 254 US patents and is an IBM Master Inventor.

Dr. Yu is a Fellow of the ACM and a Fellow of the IEEE. He is the Editor-in-Chief of *IEEE Transactions on Knowledge and Data Engineering*. He is also an associate editor of *ACM Transactions on the Internet Technology* and that of *Knowledge and Information Systems*. He is on the ICDM Steering Committee and is currently serving as the General Co-Chair of ICDM 2002.

The rest of the paper is organized as follows. In each of Sections 2 to 5, Xindong Wu presents a question, and other co-authors respond to that question with their own opinions and arguments.

2 Where Are We Now in Practical Data Mining Development?

Wu. Knowledge discovery from databases has been worked over by researchers in several disciplines including artificial intelligence and databases for over a decade. The first knowledge discovery and data mining workshop was held in Detroit, USA in August of 1989, in conjunction with the 1989 International Joint Conference on Artificial Intelligence. Data mining is an important research

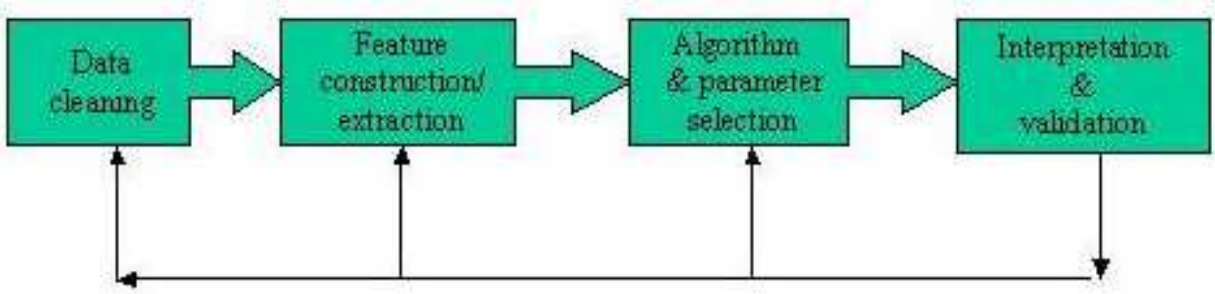

Figure 1: Data Mining Process

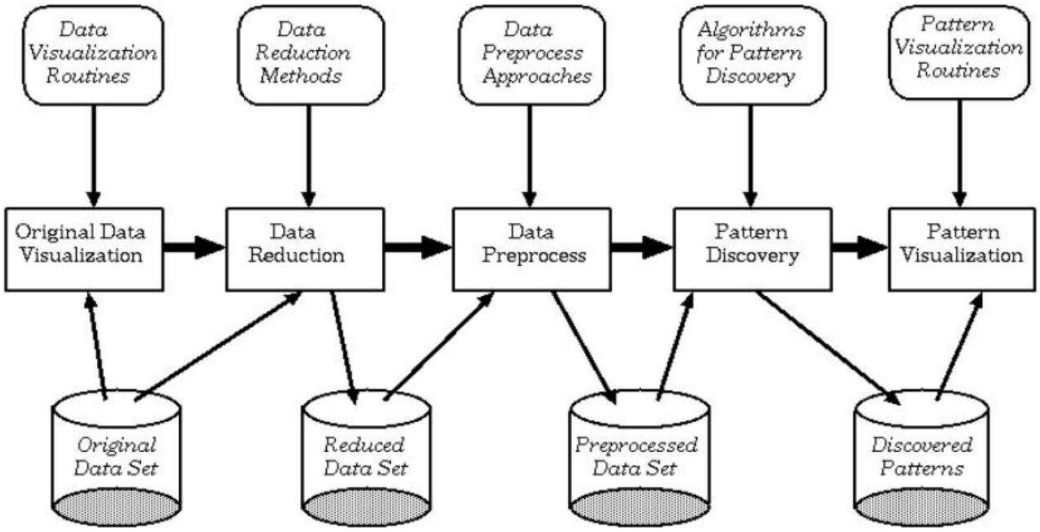

Figure 2: Data Mining: Supplementary Processes

frontier for machine learning, database technology and many other related areas. There have been many reports on successful data mining applications. In the meanwhile, the data mining field is young, and some people are still saying that it is only in its infant stage. How far is data mining from practical/industrial development?

Yu. The data mining process consists of several steps (see Figure 1): data cleaning, feature construction/extraction, algorithm and parameter selection, and interpretation and validation. Currently most research work focuses on developing new algorithms or improving the speed or accuracy of existing algorithms.

Cercone. I would like to stress the importance of supplementary processes (see Figure 2) [5] inextricably intertwined in the data mining process in order to make the data mining process more effective. Particularly I would suggest one such intertwining that uses visualization paradigms in each step of the data mining process. This forces the domain expert to become intimately involved with the process and takes advantage of his/her perceptions and skills.

There are other intertwinings that may prove advantageous as well, such as better decision tools and interfaces.

Piatetsky-Shapiro. Data mining is fundamentally an applied science. Just as the developments of probability theory by B. Pascal and statistics by R. Fisher were motivated by practical issues, rapid development of data mining applications was fueled by the business and scientific

needs to make sense of mountains of data.

Unlike fields such as neural networks and rough sets, which study a specific set of methods, data mining studies a problem – how to find useful knowledge in data and is interested in all the relevant methods that can be used.

Lin. I would like to add some facts that are less known among data miners. Machine learning, pattern recognition and some areas of AI, such as rough sets, have always engaged in some form of data mining or knowledge discovery; however, not all of them fit with the concept of modern data mining. In the 1980's, the database community extended its focus to the management of data and knowledge, including expert database systems (EDS) [6], and knowledge base management systems (KBMS) [2]. Larry Kerschberg [6] stated that the expert system component of EDS can be used to manage the rules that are extracted from databases, and Viveros (1989) [23] did exactly that.

To differentiate from classical works, I would like to propose the explicit requirement that data volumes have to be large as the marking point of the beginning of the modern era of data mining and knowledge discovery.

For conferences, the Rough Set and Database Mining Workshop at ACM CSC '95 (February, 1995) and the First International Conference on Knowledge Discovery and Data Mining (August, 1995) are probably the first two that use the terms data mining and knowledge discovery explicitly. This second conference is the beginning of the ACM SIGKDD series, and was evolved from Gregory Piatetsky-Shapiro's workshop on Knowledge Discovery in Databases (KDD '89). The first one is less known, but became one of the main conference series in the rough set community.

From the products point of view, data mining software, including industrial strength systems, is quite prosperous in numbers. Goebel and Greenwald (1999) [4] reviewed 43 products and there are many more that were not mentioned in these reviews. Unfortunately, all these software products must have some serious limitations. We have, so far, only heard about successful stories, but no major impacts on the business world have echoed. This signals that the technology of data mining may still be in its infancy.

Cercone. Just to add to Lin's excellent comments: some statisticians have worked on data analysis of large data sets for a long time but perhaps can be excluded in the vision since they do not derive results peculiar to databases or artificial intelligence. For example, Waterloo's Statistics Department have been running a Large Data Sets workshop, by invitation, for years as a means of attracting industrial collaborators.

Piatetsky-Shapiro. I disagree that data mining software is prosperous from the products point of view. It is true that there are many companies producing data mining software, but the market for stand-alone data mining software is limited. Most opportunities will be in embedded data mining software.

Kotagiri. Data mining is a vast area covering many different aspects, including clustering, classification, association rules and interestingness, time series analysis, and prediction. This area is addressed in several fields with almost no cross fertilization: statistics, machine learning, databases, control and systems theory, AI, neural computing, rough sets, and fuzzy systems.

Data mining involves many preprocessing steps. In fact many preprocessing steps are very crucial in building a useful decision system. Examples are data cleaning (which may need domain specific knowledge); and feature selection (which may involve throwing away some features, splitting some features, and combining others).

The questions are: Can we build generic preprocessing tools that can be applied in several domains, that can scale both in the number of features and the size of data? How do we provide domain specific knowledge to these tools? Can it be done in a similar way to integrity constraints in a typical database system? Answers to theses questions are crucial for making current research

in data mining to be useful in a commercial context.

Wah. Data mining has matured and is now practical for many well-defined applications, such as in predicting customer behavior in stores and in evaluating user access behavior at Web servers. These applications are generally characterized by well-defined representative data sets that can be used for training and mining. The necessary conditions for such representative data sets to be usable for data mining are that (1) they are of reasonable size, (2) they characterize approximately the behavior of the larger data warehouse, and (3) relevant conclusions can be drawn from processing them. Furthermore, tractable quantitative analysis will require that these data sets be stationary or quasi-stationary with time, the sources with correlations to these data sets be well defined, and the data be made available for analysis with little delay. These conditions are not unique to data mining but also hold for general induction-based machine learning.

3 Research vs Development in Data Mining

Wu. Research and development have different emphases. Research tends to make development more complicated than necessary. For example, most research papers in data mining deal with performance and accuracy improvement. Does a 1% increase in performance or accuracy really matter for most practical applications? The answer is probably "no!" On the other hand, development tends to simplify the research side, but puts more emphasis on the design of interfaces including interfaces with other software systems. Also, we need to distinguish two types of development; one is application oriented, and the other has a strong research component. Mining the sky, as Jim Gray addressed in his keynote talk at the conference, is an example that requires research contributions. Developing general-purpose tools for classification tasks would be an example for application oriented development.

To be more specific, how data mining research differs from practical development?

Yu. As mentioned in Section 2, research focuses on algorithmic aspects of the mining process. Practical development needs to address all steps in the mining process.

Cercone. I agree with Yu in this regard. There are implementation issues however, for example, with respect to visualization and data mining,

- Visualization spaces (1D, 2D, 3D, etc.)
- Visualization techniques
- Traversability and navigability
- Components implementation methods
- Human-machine interaction
- Window strategies (zoom in/zoom out, overview+detail, focus+context)
- Human perception
- Domain knowledge

Piatetsky-Shapiro. There are fields like pure mathematics where knowledge is important for its own sake and where researchers should not be concerned with practicality of their research. However, data mining is not such a field. I believe that *good* data mining research should be motivated by practical, real-world problems.

For example, we see many papers proposing incremental refinements in association rules algorithms, but very few papers describing how the discovered association rules are used.

Data mining researchers should be asking themselves: "Suppose I solved the problem. Now what?"

If they don't see a good application for the solution, they should be looking for a different problem. Fortunately, there are plenty of real business and scientific applications that have great need for data mining solutions, as we discuss in the next section.

Lin. The goals of research and development are quite different; one is knowledge-oriented, and the other is profit-oriented. Researchers are often excited with new ideas and new methods. However, developers, especially large scaled systems, need some degree of assurance on their reasonable returns. Such assurance may derive from two sources, market analysis and the soundness and effectiveness of the technology. The research community can possibly provide the latter, but in my opinion, it has not been done yet. We need more research towards the foundations. For example,

- Do current approaches find all possible meaningful patterns? In particular, could "invisible" association rules be mined?

"Invisible" means that the association rules are defined implicitly by the data, and more precisely, defined by derived attributes (or transformed features). Such association rules cannot be searched blindly by applying Apriori-like algorithms to the *given* data (see examples in [9]). A systematic theory of derived patterns (e.g., feature extraction and construction) is needed. Some research is initiated in [10, 11], and more work needs to be done.

- In association rule mining, the itemset frequency has been the main criterion. Is this an adequate measure?

In [11, 8], we have observed that isomorphic relations have isomorphic association rules. Since isomorphism is a syntactic notion; it is possible that two isomorphic relations may have totally different semantics. In other words, the patterns of all different semantics will be mined even though the intention is only for one application. This observation explains to some degree why there are so many association rules. We probably need more semantics oriented criteria (see [20, 21, 18]).

- I have taken the view that modern data mining distinguishes itself from classical data analysis, by explicitly stipulating that data volumes have to be large. Intuitively, it is obvious that only the patterns of large databases are interesting. But then: How large is large, and are there any explicit criteria?

In [14, 15], we have observed that, in numerical databases, both data and patterns have certain "complexity," and we required the patterns' complexity to be simpler than the data's complexity. Based on such an observation we proposed to define that "data is larger" if its complexity is larger than the patterns' complexity. The idea is adopted from the notion of randomness (no patterns) in algorithmic information theory [7]. Such notions are needed in real-time applications; we need a "hard number" to trigger the actions. Ben Wah's invited talk at ICDM 2001 [24], I believe, has implicitly taken such notions into consideration.

Kotagiri. Moving research into the commercial arena involves cost considerations. To bring a product to a commercial world requires (1) investment in research; (2) product development; and (3) marketing and product support. In general terms the investment for these phases is in the ratio of 1:10:100.

There are many other difficulties: a lot of research that takes place is incremental and may not directly yield to a product that can easily compete in the market. The size of investment can be so prohibitive that a useful concept may never be implemented as a viable product. With regard to data mining we have many of these problems. It is unclear how easily we can embed domain specific knowledge into our tools so that they can perform well. We also need these systems to learn incrementally. I would also think that users of these systems should know how these systems work so that they will have confidence in making use of these tools.

Another big problem: all we can say from a research perspective is how well a system performed on given sets of data we experimented. But we need more robust measures that can be used more meaningfully in given application domains.

Wah. Development in data mining today has focused on problems with well defined representative data sets. However, many research problems remain open. To solve problems characterized by large data warehouses, systematic techniques must be developed for abstraction and representation of relevant multi-dimensional data, numerical/syntactic/semantic analysis, multi-resolution/hierarchical analysis, and sensitivity analysis. New methods to assess the generalizability of discovered rules will also be essential.

4 What Path to Follow for Data Mining Development?

Wu. Different fields in computing have followed different paths for practical developments. The database community (in particular relational databases) and the artificial intelligence (AI) community have taken different paths. Jim Gray said in his keynote talk at ICDM 2001 that database people do not do much, but with what they do, they do a good job. This has not been the case for the AI field, partly because AI has to deal with a wider range of more complicated problems.

Should data mining concentrate on a selected number of well-defined problems such as classification and association analysis (like the database field) for development, or do we have to deal with different types of data mining problems (like AI)? If you think the data mining community has to learn from the database community by concentrating on a selected number of areas for practical development at first, what do you think are the most promising areas?

Yu. Data mining should be application driven. In addition to the e-commerce or Web applications, bioinformatics is another great area for data mining. Certainly, many techniques such as classification and clustering may be used by different applications. However, the requirements imposed by a specific application can change the nature of the mining problems and algorithms.

For example, conventional clustering is distance based, i.e., points that are close to each other on at least one subset of the dimensions are assigned to the same cluster. In bioinformatics, pattern based clustering for micro-array data is not based on distance, but on coherent patterns on a subset of the dimensions, where the expression levels of genes rise and fall synchronously in response to a set of environmental stimuli. This requires a different type of models and algorithms [25] to be devised. Another example is on document recommendation in Web mining and knowledge management. Often only documents of the positive (i.e., interested) class are available, because if one has worked on a particular task for some time, one should have accessed many related documents. A document recommender will try to recommend a similar type of documents from incoming new documents. To identify documents of the positive class from a mixed set of documents, traditional machine learning or classification techniques are not applicable, as there are no labeled documents of the non-positive class. A new type of classification scheme needs to be developed which can pick out documents of the positive class from a set of mixed documents with a training set consisting only of documents from the positive class. This is referred to as partially supervised classification [19].

By focusing on specific applications, meaningful new research problems can be formulated.

Cercone. One path that seems particularly notable is to combine the processes of data mining (a la Yu's model) with visualization and interface tools at each step in the data mining process from initial data selection, cleaning, data reduction, etc. In point form:

- Combine algorithm- and visualization-based approaches
- Attempt to visualize the entire data mining process
- Address human-machine interaction
- Take advantage of the user's perception and domain knowledge to guide the process
- Develop new visual forms to interpret the intermediate results and final patterns

Another path that has been noteworthy is the use of rough set technology in data mining, particularly in the data reduction stage.

Piatetsky-Shapiro. Investments and stock market have traditionally attracted quantitative analysts. Although in general stock market exhibits chaotic and unpredictable behavior, under some conditions it is possible to find local trends. Neural networks and genetic algorithms have been especially popular there.

Web-related and e-commerce applications generate huge amounts of data. KDD-Cup 2000, which focused on the analysis of web log, showed that it is possible to find useful knowledge in web log, but it was also very labor intensive. More automation is needed there.

Text mining and email processing are very promising applications.

Bioinformatics is indeed another excellent area. Growing popularity of microarrays and other data intensive tools make bioinformatics very attractive for data miners.

Among other good areas we should mention mobile, geographically-aware devices, mining multimedia data (such as images, video, and sound), and manufacturing.

Lin. One of the possible routes that a successful new industry can take is the following: (1) discovery of a new idea, (2) evidence of its applicability, (3) small scale systems to test the market, (4) "full" understanding of the new technology, and (5) fully scaled systems.

The developments of over forty data mining systems, none of which are in the same scale as DBMS systems, clearly indicate that we are at Step (3). To have major investments from the software industry, I believe, some foundational research that understands the power and the limit of the current technology is necessary. In the previous section, we have touched on this topic for association rule mining. For numerical mining, neural networks have been noted [16, 12] for their learning ability, and some have claimed their adaptability especially in intelligent control. Mathematically, neural networks are a methodology of "curve fitting" based on the given knowledge (in the form of activation functions). If the given knowledge does not carry the application semantics, the learning is incidental, and the learned network cannot adapt itself to the changing environments by simply adjusting its weights. Many successful experiments are due to the incidental facts that the given knowledge happens to carry the application semantics within the range of experiments.

Wah. Research and developments in data mining should focus on challenging applications that are not solved well today. Only by addressing these new applications will new and more powerful techniques be developed. Examples of such applications include the mining of remote sensing data from satellites, multi-spectral and multi-sensor data from telescopes, human genome databases, and data from financial markets.

5 Tools for Data Mining Development

Wu. With the World Wide Web's emergence as a large, distributed data repository and the realization that on-line transaction databases can be analyzed for commercial gains, data mining in large databases has attracted wide interest from both academia and the industry, and in the meanwhile, has also uncovered new challenges [22]. Data mining has its distinctive goal from related fields such as machine learning, databases and statistics, and accordingly requires distinctive tools. Which types of tools do we need for practical data mining development?

Yu. For a practical data mining tool, just having advanced algorithms alone is not sufficient. The tool needs to support the whole data mining process mentioned in Section 2, which is iterative in nature and can be application dependent. First, the tool needs to be able to interface easily with existing data of the application area. The data for e-commerce applications will be very different from the DNA or protein data for bioinformatic applications. The tool also needs to support a good user interface so that the user can interpret and validate the mining results and iteratively make feature extraction and algorithm or parameter selection during the mining process.

Cercone. I believe that applications may well determine the necessity and appropriateness of the tool to be selected. It is not just the kind of algorithm but the supporting system and interface for particular applications that will determine the usefulness of the tool.

For example, a large, complex molecular compound dataset [1] with many condition attributes, each of which may have many values, may require complex discretization algorithms to support such a situation and a method for evaluating the outcomes, be these rules induced or tables of values produced. Datasets such as those available in the UCI repository (`http://www.ics.uci.edu/~mlearn/MLSummary.html`) are relatively small and "well behaved" and may not help determine the usefulness of tools for data mining development. Large telecommunications data sets may be plagued with the same problems as the molecular compound datasets but they may be more amenable to different algorithms in the data mining engine.

I guess we could always look at the various repositories of tools and data mining systems regularly reported in KD Nuggets (http://www.kdnuggets.com/).

Lin. This topic has been well addressed by the co-authors, so I would only mention the concept of "human in the loop" and give a pointer to an implemented and marketed system [3].

Wah. Many tools for data manipulation, analysis, and visualization already exist but will need to be integrated into more user-friendly forms, involving better human-computer interfaces. Other application-specific tools will need to be developed as new and challenging applications are studied.

Wu. One area in data mining tools development that I have been pursuing since 1989 is intelligent learning database (ILDB) systems [26, 27]. An ILDB system integrates machine learning and data mining techniques with database and knowledge base technology. It starts with database technology and performs both induction and deduction. The integration of database technology, induction (from machine learning), and deduction (from knowledge-based systems) plays a key role in the construction of ILDB systems, as does the design of efficient induction and deduction algorithms.

6 Concluding Remarks

As with every field in Computing, research and development have different emphases. In data mining, we believe some of the techniques are already mature enough for practical applications, such as association analysis and classification. Research needs to meet development to get industrial support, and development needs to meet research to put researchers' ideas into viable products.

The IEEE International Conference on Data Mining provides a leading forum to explore such opportunities, and welcomes practical development experiences as well as original research results for presentation at each year's conference.

Since data mining is still a relatively new field, the co-authors of this paper (and researchers in the field in general) have different opinions on different issues discussed in this paper. Therefore, rather than providing common guidelines on how to invest in data mining research and product development, we have presented each co-author's views on each of the issues discussed in Sections 2 to 5.

To conclude the paper, we present below a few points made by the reviewers on an earlier version of this paper.

- Data mining development is linked with the development of the database field from transaction-oriented processing to on-line analysis. The path from databases for transactions to data warehouses for analysis is proceeding the path of data mining in many companies.
- Preprocessing could take around 80% of the time in a data mining application. Hence, there are aggregation and filtering operators built in modern database management systems that are expected to facilitate preprocessing. The European project MiningMart (2000 - 2003, http://www-ai.cs.uni-dortmund.de/FORSCHUNG/PROJEKTE/MININGMART/index.eng.html) is about building cases of successful preprocessing steps for further re-use.
- Modern DBMS such as Oracle 9i have provided significant operators for data mining, including preprocessing. Also Microsoft now aims at providing users with operators that decrease the burden of data inspection, data cleaning, and aggregation. IBM's data mining software deals with accessing data in a suitable way as well.
- What companies most often want is possible support in their marketing. A direct mailing action is expensive. If data mining can indeed filter out about 10% of the addresses without lowering the return, this can be a visible economic success!
- Data mining is not only different in the size of data but also in the target of learning tasks: instead of general rules with high coverage and accuracy, we are also interested in small, local patterns and negative association rules [28]. These are new learning tasks and could be in the center of data mining research and development.
- Finally, this paper contains opinions from some senior researchers in Data Mining, and we encourage people in both academia and industry to explore further the questions and opinions discussed in this paper, and get closer in development related collaborations.

Acknowledgements

The authors would like to express their appreciation to Vipin Kumar, Katharina Morik and the anonymous reviewer(s) for their constructive comments and advice on improving an earlier version of this paper.

References

[1] An, Aijun, Cercone, Nick, and Huang, Xianji, A Case-Study for Learning from Imbalanced Sets, *CSCSI/SCEIO 2001*, Ottawa, Canada, 2001, pp. 1-15.

[2] Brodie, M., and Mylopoulou, J., *On Knowledge Base Management Systems*, Springer-Verlag, 1986.

[3] Chiang, I., and Lin T.Y., Index Miner. *Proceedings of the International Conference on Computer Software and Applications*, Chicago, October 8-12, 2001. pp. 613-614

[4] Goebel, M., and Greenwald, L., A Survey of Data Mining and Knowledge Discovery Tools, *ACM SIGKDD Explorations*, Vol 1, No 1, June 1999, pp. 20-33.

[5] Han, J., and Cercone, N., CVis: An Interactive Visualization System for Rule Induction, 13th CSCSI/SCEIO Conference, *Lecture Notes in Artificial Intelligence 1822, Springer*, Montreal, PQ, 2000, pp. 214-226. [This paper won the best paper award].

[6] Kerschberg, L. (ed.), *Expert Database Systems: Proceedings from the First International Conference* (October 24-27, 1984), Benjamin/Cummings, Menlo Park, CA, 1986.

[7] Li, M., and Vitanyi, P., *Introduction to Komogorov Complexity and Its Applications*, Springer Verlag, 1997.

[8] Lin, T.Y., Feature Completion, *Communications of IICM (the Institute of Information and Computing Machinery, Taiwan)*, Vol 5, No. 2, May 2002, pp. 57-62. (This is the Proceedings of the Workshop "Towards the Foundation on Data Mining" at PAKDD 2002, May 6, 2002.)

[9] Lin, T.Y., Issues in Data Mining, *Proceedings of 26th IEEE International Conference on Computer Software and Applications*, Oxford, UK, Aug 26-29, 2002.

[10] Lin, T.Y., Attributes Transformations for Data Mining: Theoretical Explorations, *International Journal of Intelligent Systems*, 2002 (*to appear*).

[11] Lin, T.Y., Data Mining on Derived Attributes - A Granular and (Non-Standard) Rough Computing Approach, *Proceedings of the 2002 International Conference on Rough Sets and Current Trends for Computing* (To appear in *Lecture Notes in AI*), Springer Verlag.

[12] Lin, T.Y., Neural Networks, Qualitative-Fuzzy Logic and Granular Adaptive Systems, *Proceedings of the 2002 World Congress of Computational Intelligence*, Honolulu, Hawaii, May 12-17, 2002, 566-571.

[13] Lin, T.Y. The Lattice Structure of Database and Mining High Level Rules, *Proceedings of Workshop on Data Mining and E-Organizations, International Conference on Computer Software and Applications*, Chicago, 2001, October 8-12, 2001. Also, to appear as "Feature Transformations and Structure of Attributes" in *Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, B. Dasarathy (ed), Proceeding of SPIE Vol 4730, Orlando,Fl, April 1-5, 2002.

[14] Lin, T.Y. Discovering Patterns in Numerical Sequences Using Rough Set Theory, *Proceeding of the Third World Multi-Conference on Systemics, Cybernetics and Informatics*, Orlando, Florida, July 31 - August 4, 1999, Vol 5, 568-572.

[15] Lin, T.Y. Patterns in Numerical Data: Practical Approximations to Kolmogorov Complexity, *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing (Lecture Notes in Artificial Intelligence No 1711)*, Zhong, Skowron, Ohsuga (eds), Springer-Verlag, 1999, 509-513.

[16] Lin, T.Y., The Power and Limit of Neural Networks, *Proceedings of the 1996 Engineering Systems Design and Analysis Conference*, Montpellier, France, July 1-4, 1996, Vol. 7, pp. 49-53.

[17] Lin, T.Y., and Tremba, J., Attribute Transformations on Numerical Databases: Applications to Stock Market Data, *Methodologies for Knowledge Discovery and Data Mining (Lecture Notes in Artificial Intelligence No 1805)*, T. Terano, H. Liu and A. Chen (eds), Springer-Verlag, 2000, pp. 181-192.

[18] Lin, T.Y., Zhong, N., Duong, J. and Ohsuga, S., Frameworks for Mining Binary Relations in Data, *Rough Sets and Current Trends in Computing (Lecture Notes in Artificial Intelligence 1424)*, A. Skoworn and L. Polkowski (eds), Springer-Verlag, 1998, pp. 387-393.

[19] Liu, B., Lee, W.S., Yu, P.S. and Li, X., Partially Supervised Classification of Text Documents, *Proc. 19th Intl. Conf. on Machine Learning*, Sydney, Australia, July 2002, 387-394.

[20] Louie, E, and Lin., T.Y., Semantics Oriented Association Rules, *Proceedings of the 2002 World Congress of Computational Intelligence*, Honolulu, Hawaii, May 12-17, 2002, 956-961.

[21] Ng, R., Lakshmanan, L.V.S., Han, J. and Pang, A. Exploratory Mining and Pruning Optimizations of Constrained Associations Rules, *Proceedings of the 1998 ACM-SIGMOD Conference on Management of Data*, 1998, pp. 13-24.

[22] Ramakrishnan, N. and Grama, A.Y., Data Mining: From Serendipity to Science. *IEEE Computer*, 32, 1999, 8: 34–37.

[23] Viveros, M., Extraction of Knowledge from Databases, *Thesis*, California State University at Northridge, 1989

[24] Wah, Benjamin W. and Qian, Minglun, Constrained Formulations and Algorithms for Stock-Price Predictions Using Recurrent FIR Neural Networks, *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002)*, Edmonton, Alberta, Canada, 2002.

[25] Wang, H., Yang, J., Wang, W., and Yu, P.S., Clustering by Pattern Similarity in Large Data Sets, *Proc. ACM SIGMOD Conference*, Madison, WI, June 2002.

[26] Wu, X., *Knowledge Acquisition from Databases.* Ablex Publishing Corp., U.S.A., 1995.

[27] Wu, X., Building Intelligent Learning Database Systems. *AI Magazine*, 21, 2000, 3: 59-65.

[28] Wu, X., Zhang, C. and Zhang, S., Mining Both Positive and Negative Association Rules, *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, The University of New South Wales, Sydney, Australia, 8-12 July 2002, 658-665.