Research Challenges for Data Mining in Science and Engineering*

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

ABSTRACT

With the rapid development of computer and information technology in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in gigantic storage devices or flowing into and out of the system in the form of data streams. Moreover, such data has been made widely available, e.g., via the Internet. Such tremendous amount of data, in the order of tera- to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new, data-intensive methods to conduct research in science and engineering.

In this paper, we discuss the research challenges in science and engineering, from the data mining perspective, with a focus on the following issues: (1) *information network analysis*, (2) *discovery, usage, and understanding of patterns and knowledge*, (3) *stream data mining*, (4) *mining moving object data, RFID data, and data from sensor networks*, (5) *spatiotemporal and multimedia data mining*, (6) *mining text, Web, and other unstructured data*, (7) *data cube-oriented multidimensional online analytical mining*, (8) *visual data mining*, and (9) *data mining by integration of sophisticated scientific and engineering domain knowledge*.

1. INTRODUCTION

It has been popularly recognized that the rapid development of computer and information technology in the last twenty years has fundamentally changed almost every field in science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for the development of new, data-intensive methods to conduct research in science and engineering. Thus the new terms like, *data science* [13] or *data engineering*, can be used to best characterize the data-intensive nature of today's science and engineering.

Besides the further development of database methods to efficiently store and manage peta-bytes of data online, making these archives easily and safely accessible via the Internet and/or a computing grid, another essential task is to develop powerful data mining tools to analyze such data. Thus, there is no wonder that data mining has also entered on to the center stage in science and engineering.

Data mining, as the confluence of multiple intertwined disciplines, including *statistics, machine learning, pattern recognition, database systems, information retrieval*, *World-Wide Web*, *visualization*, and *many application domains*, has made great progress in the past decade [15]. To ensure that the advances of data mining research and technology will effectively benefit the progress of science and engineering, it is important to examine the challenges on data mining posed in data-intensive science and engineering and explore how to further develop the technology to facilitate new discoveries and advances in science and engineering.

2. MAJOR RESEARCH CHALLENGES

In this section, we will examine the major challenges raised in science and engineering from the data mining perspective, and examine the promising research directions.

2.1 Information network analysis

With the development of Google and other effective web search engines, information network analysis has become an important research frontier, with broad applications, such as social network analysis, web community discovery, terrorist network mining, computer network analysis, and network intrusion detection. However, information network research should go beyond explicitly formed, homogeneous networks (*e.g.*, web page links, computer networks, and terrorist e-connection networks) and delve deeply into *implicitly formed*, *heterogeneous*, and *multidimensional* information networks. Science and engineering provide us with rich opportunities on exploration of networks in this direction.

There are a lot of massive natural, technical, social, and information networks in science and engineering applications, such as gene/protein/microarray networks in biology, highway transportation networks in civil engineering, topic/theme-author-publication-citation networks in library science, wireless telecommunication networks among commanders, soldiers and supply lines in a battle field. In such information networks, each node or link in a network con-

*The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678 and BDI-05-15813. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

tains *valuable, multidimensional information*, such as textual contents, geographic information, traffic flow, and other properties. Moreover, such networks could be highly *dynamic, evolving, and inter-dependent*.

Although a single link in a network could be noisy, unreliable, and sometimes misleading, valuable knowledge can be mined reliably among a large number of links in a massive information network. Our recent studies on information networks show that the power of such links in massive information networks should not be underestimated. They can be used for predictive modeling across multiple relations [30], for user-guided clustering across multiple relations [31], for effective link-based clustering [16, 32], for distinguishing different objects with identical names [33] and for solving the veracity problem, *i.e.*, finding reliable facts among multiple conflicting web information providers [34]. The power of such links should be thoroughly explored in many scientific domains, such as in protein network analysis in biology and in the analysis of networks of research publications in library science as well as in each science/engineering discipline.

Another important direction in information network analysis is to treat information networks as graphs and further develop graph mining methods [8, 29]. Recent progress on graph mining and its associated structural pattern-based classification and clustering, graph and graph containment indexing, and similarity search will play an important role in information network analysis. Moreover, since information networks often form huge, multidimensional heterogeneous graphs, mining noisy, approximate, and heterogeneous subgraphs based on different applications for the construction of application-specific networks with sophisticated structures will help information network analysis substantially. The use of the power law distribution of many information networks and the rules on density evolution of information networks will help reduce computational complexity and enhance to power of network analysis. Finally, the study of link analysis, heterogeneous data integration, user-guided clustering, user-based network construction, will provide essential methodology for the in-depth study in this direction.

2.2 Discovery, understanding, and usage of patterns and knowledge

Scientific and engineering applications often handle massive data of high dimensionality. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers.

Frequent pattern mining has been a focused theme in data mining research for over a decade [14]. Abundant literature has been dedicated to this research, and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structural pattern mining, correlation mining, associative classification, and frequent-pattern-based clustering, as well as their broad applications.

Recently, studies have proceeded to scalable methods for mining colossal patterns [35] where the size of the patterns could be rather large so that the step-by-step growth using an Apriori-like approach does not work, methods for pattern compression, extraction of high-quality top-k patterns [28], and understanding patterns by context analysis and generation of semantic annotations [22]. Moreover, frequent patterns have been used for effective classification by top-k rule generation for long patterns and discriminative frequent pattern analysis [7]. Frequent patterns have also been used for clustering of high-dimensional biological data [26]. Moreover, much research has been done on effective sequential and structural pattern mining methods and the exploration of their applications [14, 8].

The promotion of effective application of pattern analysis methods in scientific and engineering applications is an important task in data mining. Moreover, it is important to further develop efficient methods for mining long, approximate, compressed, and sophisticated patterns for advanced applications, such as mining biological sequences and networks and mining patterns related to scientific and engineering processes. Furthermore, the exploration of mined patterns for classification, clustering, correlation analysis, and pattern understanding will still be interesting topics in research.

2.3 Stream data mining

Stream data refers to the data that flows into and out of the system like streams. Stream data is usually in vast volume, changing dynamically, possibly infinite, and containing multi-dimensional features. Typical examples of such data include audio and video recording of scientific and engineering processes, computer network information flow, web click streams, and satellite data flow. Such data cannot be handled by traditional database systems, and moreover, most systems may only be able to read a data stream once in sequential order. This poses great challenges on effective mining of stream data.

With substantial research [1], progress has bee made on efficient methods for mining frequent patterns in data streams, multidimensional analysis of stream data (such as construction of stream cubes), stream data classification, stream clustering, stream outlier analysis, rare event detection [10], and so on. The general philosophy is to develop single-scan algorithms to collect information about stream data in tilted time windows, exploring micro-clustering, limited aggregation, and approximation. For skewed distribution of stream data, it is recommended to explore biased selective sampling and robust emsemble methods in model construction [10].

Stream data is often encountered in science and engineering applications. It is important to explore stream data mining in such applications and develop application-specific methods, *e.g.*, real-time anomaly detection in computer network analysis, in electric power grid supervision, in weather modeling, in engineering and security surveillance, and other stream data applications.

2.4 Mining moving object data, RFID data, and data from sensor networks

With the popularity of sensor networks, GPS, cellular phones, other mobile devices, and RFID technology, tremendous amount of moving object data has been collected, calling for effective analysis. This is especially true in many scientific, engineering, business and homeland security applications.

Interesting research has been conducted on warehousing and mining RFID data sets [11], detection of strange moving objects [19], clustering trajectory data [17], and mining traffic data for route planning [12]. However, this is still a young field with many research issues to be explored on mining moving object data, RFID data, and data from sen-

sor networks. For example, how to explore correlation and regularity to clean noisy sensor network and RFID data, how to integrate and construct data warehouses for such data, how to perform scalable mining for peta-byte RFID data, how to find strange moving objects, how to classify multidimensional trajectory data, and so on. With time, location, moving direction, speed, as well as multidimensional semantics of moving object data, likely multi-dimensional data mining will play an essential role in this study.

2.5 Spatial, temporal, spatiotemporal, and multimedia data mining

Scientific and engineering data is usually related to space, time, and in multimedia modes (*e.g.*, containing color, image, audio, and video). With the popularity of digital photos, audio DVDs, videos, YouTube, web-based map services, weather services, satellite images, digital earth, and many other forms of multimedia, spatial, and spatiotemporal data, mining spatial, temporal, spatiotemporal, and multimedia data will become increasingly popular, with far-reaching implications [23, 24]. For example, mining satellite images may help detect forest fire, find unusual phenomena on earth, predict hurricane landing site, discover weather patterns, and outline global warming trends.

Research in this domain needs the confluence of multiple disciplines including image processing, pattern recognition, geographic information systems, parallel processing, and statistical data analysis. Automatic categorization of images and videos, classification of spatiotemporal data, finding frequent/sequential patterns and outliers, spatial collocation analysis, and many other tasks have been studied popularly. With the mounting of such data, the development of scalable analysis methods and new data mining functions will be an important research frontier for years to come.

2.6 Mining text, Web, and other unstructured data

Web is the common place for scientists and engineers to publish their data, share their observations and experiences, and exchange their ideas. There is a tremendous amount of scientific and engineering data on the web. For example, in biology and bioinformatics research, there are GenBank, ProteinBank, GO, PubMed, and many other biological or biomedical information repositories available on the Web. Therefore, the Web has become the ultimate information access and processing platform, housing not only billions of link-accessed "pages", containing textual data, multimedia data, and linkages, on the surface Web, but also query-accessible "databases" on the deep Web. With the advent of Web 2.0, there is an increasing amount of dynamic "workflow" emerging. With its penetrating deeply into our daily life and evolving into unlimited dynamic applications, the Web is central in our information infrastructure. Its virtually unlimited scope and scale render immense opportunities for data mining.

Text mining and information extraction have been applied not only to Web mining but also to the analysis of other kinds of semi-structured and unstructured information, such as digital libraries, biological information systems, research literature analysis systems, computer-aided design and instruction, and office automation systems.

There are lots of research issues in this domain [4, 20], which takes collaborative efforts of multiple disciplines, including information retrieval, databases, data mining, natural language processing, and machine learning. For many scientific and engineering applications, the data is somewhat structured and semi-structured, with designated fields for text and multimedia data. Thus it is possible to mine and build relatively structured web repositories. Some promising research topics include heterogeneous information integration, information extraction, personalized information agents, application-specific partial Web construction and mining, in-depth Web semantics analysis, development of scientific and engineering domain-specific semantic Webs, and turning Web into relatively structured information-base.

2.7 Data cube-oriented multidimensional online analytical mining

Scientific and engineering datasets are usually high-dimensional in nature. Viewing and mining data in multidimensional space will substantially increase the power and flexibility of data analysis. Data cube computation and OLAP (online analytical processing) technologies developed in data warehouse have substantially increased the power of multidimensional analysis of large datasets. Besides traditional data cubes, there are recent studies on construction of regression cubes [6], prediction cubes [5], and other scalable high-dimensional data analysis methods [18]. Such multidimensional, especially high-dimensional, analysis tools will ensure data can be analyzed in hierarchical, multidimensional structures efficiently and flexibly at user's finger tips. This leads to the integration of online analytical processing with data mining, *i.e.*, OLAP mining.

We believe that OLAP mining will substantially enhance the power and flexibility of data analysis and lead to the construction of easy-to-use tools for the analysis of massive data with hierarchical structures in multidimensional space. It is a promising research field for developing effective tools and scalable methods for exploratory-based scientific and engineering data mining.

2.8 Visual data mining

A picture is worth a thousand words. There have been numerous data visualization tools for visualizing various kinds of data sets in massive amount and of multidimensional space [25]. Besides popular bar charts, pie charts, curves, histograms, quantile plots, quantitle-quantile plots, boxplots, scatter plots, there are also many visualization tools using geometric (*e.g.*, dimension stacking, parallel coordinates), hierarchical (*e.g.*, treemap), and icon-based (*e.g.*, Chernoff faces and stick figures) techniques. Moreover, there are methods for visualizing sequences, time-series data, phylogenetic trees, graphs, networks, web, as well as various kinds of patterns and knowledge (*e.g.*, decision-trees, association rules, clusters and outliers) [9]. There are also visual data mining tools that may facilitate interactive mining based on user's judgement of intermediate data mining results [2]. Recently, we have developed a DataScope system that maps relational data into 2-D maps so that multidimensional relational data can be browsed in Google map's way [27].

We believe that visual data mining is appealing to scientists and engineers because they often have good understanding of their data, can use their knowledge to interpret their data and patterns with the help of visualization tools, and interact with the system for deeper and more effective mining. Tools should be developed for mapping data

and knowledge into appealing and easy-to-understand visual forms, and for interactive browsing, drilling, scrolling, and zooming data and patterns to facilitate user exploration. Finally, for visualization of large amount of data, parallel processing and high-performance visualization tools should be investigated to ensure high performance and fast response.

2.9 Domain-specific data mining: Data mining by integration of sophisticated scientific and engineering domain knowledge

Besides general data mining methods and tools for science and engineering, each scientific or engineering discipline has its own data sets and special mining requirements, some could be rather different from the general ones. Therefore, in-depth investigation of each problem domain and development of dedicated analysis tools are essential to the success of data mining in this domain. Here we examine two problem domains: biology and software engineering.

Biological data mining

The fast progress of biomedical and bioinformatics research has led to the accumulation and publication (on the web) of vast amount of biological and bioinformatics data. However, the analysis of such data poses much greater challenges than traditional data analysis methods [3]. For example, genes and proteins are gigantic in size (*e.g.*, a DNA sequence could be in billions of base pairs), very sophisticated in function, and the patterns of their interactions are largely unknown. Thus it is a fertile field to develop sophisticated data mining methods for in-depth bioinformatics research. We believe substantial research is badly needed to produce powerful mining tools in many biological and bioinformatics subfields, including comparative genomics, evolution and phylogeny, biological data cleaning and integration, biological sequence analysis, biological network analysis, biological image analysis, biological literature analysis (*e.g.*, PubMed), and systems biology. From this point view, data mining is still very young with respect to biology and bioinformatics applications. Substantial research should be conducted to cover the vast spectrum of data analysis tasks.

Data mining for software engineering

Software program executions potentially (*e.g.*, when program execution traces are turned on) generate huge amounts of data. However, such data sets are rather different from the datasets generated from the nature or collected from video cameras since they represent the executions of program logics coded by human programmers. It is important to mine such data to monitor program execution status, improve system performance, isolate software bugs, detect software plagiarism, analyze programming system faults, and recognize system malfunctions.

Data mining for software engineering can be partitioned into static analysis and dynamic/stream analysis, based on whether the system can collect traces beforehand for post-analysis or it must react at real time to handle online data. Different methods have been developed in this domain by integration and extension of the methods developed in machine learning, data mining, pattern recognition, and statistics. For example, statistical analysis such as hypothesis testing) approach [21] can be performed on program execution traces to isolate the locations of bugs which distinguish program success runs from failing runs. Despite of its limited success, it is still a rich domain for data miners to research and further develop sophisticated, scalable, and real-time data mining methods.

3. CONCLUSIONS

Science and engineering are fertile lands for data mining. In the last two decades, science and engineering have evolved to a stage that gigantic amounts of data are constantly being generates and collected, and data mining and knowledge discovery becomes the essential scientific discovery process. We have proceeded to the era of *data science* and *data engineering*.

In this paper, we have examined a few important research challenges in science and engineering data mining. There are still several interesting research issues not covered in this short abstract. One such issue is the development of *invisible data mining* functionality for science and engineering which builds data mining functions as an invisible process in the system (*e.g.*, rank the results based on the relevance and some sophisticated, preprocessed evaluation functions) so that users may not even sense that data mining has been performed beforehand or is being performed and their browsing and mouse clicking are simply using the results of or further exploring of data mining. Another research issue is *privacy-preserving data mining* that aims to performing effective data mining without disclosure of private or sensitive information to outsiders. Finally, *knowledge-guided intelligent human computer interaction* based on the knowledge extracted from data could be another interesting issue for future research.

4. REFERENCES

[1] C. C. Aggarwal. *Data Streams: Models and Algorithms*. Kluwer Academic, 2006.

[2] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: An interactive approach to decision tree construction. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 392–396, San Diego, CA, Aug. 1999.

[3] P. Bajcsy, J. Han, L. Liu, and J. Yang. Survey of bio-data analysis from data mining perspective. In Jason T. L. Wang, Mohammed J. Zaki, Hannu T. T. Toivonen, and Dennis Shasha, editors, *Data Mining in Bioinformatics*, pages 9–39. Springer Verlag, 2004.

[4] S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data*. Morgan Kaufmann, 2002.

[5] B.-C. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction cubes. In *Proc. 2005 Int. Conf. Very Large Data Bases (VLDB'05)*, pages 982–993, Trondheim, Norway, Aug. 2005.

[6] Y. Chen, G. Dong, J. Han, J. Pei, B. W. Wah, and J. Wang. Regression cubes with lossless compression and aggregation. *IEEE Trans. Knowledge and Data Engineering*, 18:1585–1599, 2006.

[7] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[8] D. J. Cook and L. B. Holder. *Mining Graph Data*. John Wiley & Sons, 2007.

[9] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge*

Discovery. Morgan Kaufmann, 2001.

[10] J. Gao, W. Fan, J. Han, and P. S. Yu. A general framework for mining concept-drifting data streams with skewed distributions. In *Proc. 2007 SIAM Int. Conf. Data Mining (SDM'07)*, Minneapolis, MN, April 2007.

[11] H. Gonzalez, J. Han, X. Li, and D. Klabjan. Warehousing and analysis of massive RFID data sets. In *Proc. 2006 Int. Conf. Data Engineering (ICDE'06)*, page 83, Atlanta, Georgia, April 2006.

[12] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag. Adaptive fastest path computation on a road network: A traffic mining approach. In *Proc. 2007 Int. Conf. on Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[13] J. Gray and A. Szalay. The world wide telescope: An archetype for online science. *Comm. ACM*, 45:50–54, Nov. 2002.

[14] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15:55–86, 2007.

[15] J. Han and M. Kamber. *Data Mining: Concepts and Techniques (2nd ed.)*. Morgan Kaufmann, 2006.

[16] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proc. 2002 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 538–543, Edmonton, Canada, July 2002.

[17] J.-G. Lee, J. Han, and K. Whang. Clustering trajectory data. In *Proc. 2007 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'07)*, Beijing, China, June 2007.

[18] X. Li and J. Han. Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proc. 2007 Int. Conf. on Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[19] X. Li, J. Han, S. Kim, and H. Gonzalez. Roam: Rule- and motif-based anomaly detection in massive moving object data sets. In *Proc. 2007 SIAM Int. Conf. Data Mining (SDM'07)*, Minneapolis, MN, April 2007.

[20] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.

[21] C. Liu, L. Fei, X. Yan, J. Han, and S. P. Midkiff. Statistical debugging: A hypothesis testing-based approach. *IEEE Trans. Software Engineering*, 32:831–848, 2006.

[22] Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai. Generating semantic annotations for frequent patterns with context analysis. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 337–346, Philadelphia, PA, Aug. 2006.

[23] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[24] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[25] E. R. Tufte. *The Visual Display of Quantitative Information (2nd ed.)*. Graphics Press, 2001.

[26] H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proc. 2002 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'02)*, pages 418–427, Madison, WI, June 2002.

[27] T. Wu, X. Li, D. Xin, J. Han, J. Lee, and R. Redder. Datascope: Viewing database contents in google maps' way. In *Proc. 2007 Int. Conf. Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[28] D. Xin, H. Cheng, X. Yan, and J. Han. Extracting redundancy-aware top-k patterns. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 444–453, Philadelphia, PA, Aug. 2006.

[29] X. Yan and J. Han. Discovery of frequent substructures. In *D. Cook and L. Holder (ed.), Mining Graph Data*, pages 99–115, John Wiley Sons, 2007.

[30] X. Yin, J. Han, J. Yang, and P. S. Yu. Efficient classification across multiple database relations: A crossmine approach. *IEEE Trans. Knowledge and Data Engineering*, 18:770–783, 2006.

[31] X. Yin, J. Han, and P. S. Yu. Cross-relational clustering with user's guidance. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'05)*, pages 344–353, Chicago, IL, Aug. 2005.

[32] X. Yin, J. Han, and P. S. Yu. Linkclus: Efficient clustering via heterogeneous semantic links. In *Proc. 2006 Int. Conf. on Very Large Data Bases (VLDB'06)*, Seoul, Korea, Sept. 2006.

[33] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[34] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)*, San Jose, CA, Aug. 2007.

[35] F. Zhu, X. Yan, J. Han, P. S. Yu, and H. Cheng. Mining colossal frequent patterns by core pattern fusion. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.