CHALLENGES FOR DATA MINING

Laxmi Choudhary*

ABSTRACT

With the fast development of computer and information technology in the last many years, 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.

Keywords: *Data Mining, Data Engineering, Knowledge Discovery.*

* Banasthali University, Jaipur.

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, there is no wonder that data mining has also stepped on to the center stage in science and engineering. Data mining, as the continuance 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. 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 several major challenges raised in science and engineering from the data mining perspective, and point out some 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. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to inappropriate conclusions about the data. In fact, object linkage is knowledge that should be exploited. 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. 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.

The most well known link mining task is that of link-based object ranking (LBR), which is a primary focus of the link analysis community. The objective of LBR is to exploit the link structure of a graph to order or prioritize the set of objects within the graph. Since the introduction of the most notable approaches, PageRank and HITS, many variations have been developed to rank one type or multiple types of objects in the graph. Also, the link-based object classification (LBC) problem has been studied. The task is to predict the class label for each object. The discerning feature of LBC that makes it different from traditional classification is that in many cases, the labels of related objects tend to be correlated. The challenge is to design algorithms for collective classification that exploit such correlations and jointly infer the categorical values associated with the objects in the graph. Another link-related task is entity resolution, which involves identifying the set of objects in a domain. The goal of entity resolution is to determine which references in the data refer to the same real-world entity. Examples of this problem arise in databases (de-duplication, data integration), natural language processing (co-reference resolution, object consolidation), personal information management, and other fields. Recently, there has been significant interest in the use of links for improved entity resolution.

One line of work attempts to find frequent subgraphs, and some other lines of work are on efficient subgraph generation and compression-based heuristic search. 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. Finally, the studies of link analysis, heterogeneous data integration, user-guided clustering, and user-based network construction will provide essential methodology for the in-depth study in this direction.

Many domains of interest today are best described as a network of interrelated heterogeneous objects. As future work, link mining may focus on the integration of link mining algorithms for a spectrum of knowledge discovery tasks.

2.2 Discovery, Understanding, and Usage of Patterns and Knowledge

Scientific and engineering applications often handle massive data of high dimensionality. The goal of pattern mining is to find item-sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers.

There are also various proposals on reduction of such a huge set, including closed patterns, maximal patterns, approximate patterns, condensed pattern bases, representative patterns, clustered patterns, and discriminative frequent patterns. We also need mechanisms for deep understanding and interpretation of patterns, e.g., semantic annotation for frequent patterns, and contextual analysis of frequent patterns. The main research work on pattern analysis has been focused on pattern composition (e.g., the set of items in item-set patterns) and frequency. A contextual analysis of frequent patterns over the structural information can help respond questions. The deep understanding of frequent patterns is essential to improve the interpretability and the usability of frequent patterns. Besides studies on transaction datasets, much research has been done on effective sequential and structural pattern mining methods and the exploration of their applications. 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. First, the techniques to summarize the whole or part of the data streams are studied, which is the basis for stream data mining. Such techniques include sampling, load shedding and sketching techniques, synopsis data structures, stream cubing, and clustering. The focus of stream pattern analysis is to approximate the frequency counts for infinite stream data. Algorithms have been developed to count frequency using tilted windows based on the fact that users are more interested in the most recent transactions; approximate frequency counting based on previous historical data to calculate the frequent patterns incrementally and track the most frequent k items in the

continuously arriving data. Initial studies on stream clustering concentrated on extending K-means and K-median algorithms to stream environment. The main idea behind the developed algorithms is that the cluster centers and weights are updated after examining one transaction or a batch of transactions, whereas the constraints on memory and time complexity are satisfied by limiting the number of centers.

Projected clustering can also be performing for high dimensional data streams. The focus of stream classification of data streams is first on how to efficiently update the classification model when data continuously flow in streams. Stream data is often encountered in science and engineering 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, tremensdous 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. Sensor networks are finding increasing number of applications in many domains, including battle fields, smart buildings, and even the human body. Moreover, sensors must process a continuous (possibly fast) stream of data. Data mining in wireless sensor networks (WSNs) is a challenging area, as algorithms need to work in extremely demanding and constrained environment of sensor networks (such as limited energy, storage, computational power, and bandwidth). WSNs also require highly decentralized algorithms. In designing algorithms for sensor networks, it is imperative to keep in mind that power consumption has to be minimized. Even gathering the distributed sensor data in a single site could be expensive in terms of battery power consumed, some attempts have been made towards making the data collection task energy efficient and balance the energy-quality trade-offs. Clustering the nodes of the sensor networks is an important optimization problem. Nodes that are clustered together can easily communicate with each other, which can be applied to energy optimization and developing optimal algorithms for clustering sensor nodes. Other works in this field include identification of rare events or anomalies, finding frequent itemsets, and data preprocessing in sensor networks. These movement data, including RFID data, object trajectories, anonymous aggregate data such as the one generated by many road sensors, contain rich information.

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. 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. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial data sets. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Interesting research topics in this field include prediction of events at particular geographic locations, detecting spatial outliers whose no-spatial attributes are extreme relative to its neighbors, finding co-location patterns where instances containing the patterns often located in close geographic proximity, and grouping a set of spatial objects into clusters. Future research is needed to compare the difference and similarity between classical data mining and spatial data mining techniques, model semantically rich spatial properties other than neighborhood relationships, design effective statistical methods to interpret the mined spatial patterns, investigate proper measures for location prediction to improve spatial accuracy and facilitate visualization of spatial relationships by representing both spatial and non-spatial features. With the generated features, mining can be carried out using data mining techniques to discover significant patterns. These resulting patterns are then evaluated and interpreted in order to obtain the final applications knowledge. Numerous methodologies have been developed and many applications have been investigated, including organizing multimedia data indexing and retrieval, extracting representative features from raw multimedia data before the mining process and integrating features obtained from multiple modalities.

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 theWeb. 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. Text summarization helps users figure out whether a lengthy document meets their needs and is worth reading. With large-volume texts, text-summarization software processes and summarizes the document in almost no time. The key to summarization is reducing the length and detail of a document while retaining its main points and overall meaning. Categorization involves identifying the main themes of a document.

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World-Wide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is an automatic process that goes beyond keyword extraction to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. The text content is the most widely researched area. The technologies that are normally used in web content mining are natural language processing, information retrieval, and text mining. 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. In, the issues of anomaly detection in multi-dimensional time-series data are examinedRecent study on sampling cubes discuss about the desirability of OLAP over sampling data, which may not represent the full data in the population. The proposed sampling cube framework could efficiently calculate confidence intervals for any multidimensional query and uses the OLAP structure to group similar segments to increase sampling size when needed. Further, to handle high dimensional data, a Sampling Cube Shell method is proposed to effectively reduce the storage requirement while still preserving query result quality. Such multi-dimensional, 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. Some efforts have been devoted along this direction, but grand challenge still exits when one needs to explore the large space of choices to find interesting patterns and trends. 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. 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 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. There are also visual data mining tools that may facilitate interactive mining based on user's judgement of intermediate data mining results. Recently, we have developed a Data scope system that maps relational data into 2-D maps so that multidimensional relational data can be browsed in Google map's way.

Most data analysts use visualization as part of a process sandwich strategy of interleaving mining and visualization to reach a goal, an approach commonly identified in many research works on applications and techniques for visual data mining. Usually, the analytical mining techniques themselves do not rely on visualization. Most of the papers describing visual data mining approaches and applications found in the literature fall into two categories: either they use visual data exploration systems or techniques to support a knowledge extraction goal or a specific mining task, or they use visualization to display the results of a mining algorithm, such as a clustering process or a classifier, and thus enhance user comprehension of the results. Many mining techniques involve different mathematical steps that require user intervention. Some of these can be quite complex and visualization can support the decision processes involved in making such interventions. From this viewpoint, a visual data mining technique is not just a visualization technique being applied to exploit data in some phases of

an analytical mining process, but a data mining algorithm in which visualization plays a major role.

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.

2.9.1 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. 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. 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.

2.9.2 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. 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 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. 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 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. Edoardo M. Airoldi and Kathleen M. Carley. Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings. SIGKDD Explor. Newsl. 7(2):13{22, 2005.
2. R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data Warehouses. In Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02), pages 586{597, Hong Kong, China, Aug. 2002.
3. M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: Aninteractive approach to decision tree construction. In Proc. 1999 Int. Conf. KnowledgeDiscovery and Data Mining (KDD'99), pages 392{396, San Diego, CA, Aug. 1999.
4. C. C. Aggarwal. Data Streams: Models and Algorithms. Kluwer Academic, 2006.
5. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03), pages 81{92, Berlin, Germany, Sept. 2003.
6. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for projected clustering of High dimensional data streams. In Proc. 2004 Int. Conf. Very Large Data Bases (VLDB'04), pages 852{863, Toronto, Canada, Aug. 2004.

7. James Allan. Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell, MA, USA, 2002.
8. B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02), pages 1{16, Madison, WI, June 2002.
9. M. W. Berry. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2003.
10. I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In Proc. 2006 SIAM Int. Conf. Data Mining (SDM'06), Bethesda, MD, April 2006.
11. 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, DataMining in Bioinformatics, pages 9{39. Springer Verlag, 2004.