Journal of Advances in Science and Technology

Vol. IV, No. VII, November-2012, ISSN 2230-9659

A STUDY OF DATA MINING IN SCIENCE AND ENGINEERING

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, No. VII, November-2012, ISSN 2230-9659

A Study of Data Mining in Science and Engineering

Ikramuddin Dyer[1] Dr. Pardeep Goal[2]

[1]Research Scholar, Singhania University, Rajasthan

[2]Associate Prof., M.M. College, Faitiabad

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 following 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.

----------♦----------

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 or data-intensive 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 stepped on to the center stage in science and engineering. Data mining, as the consequence 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 [HK06]. 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.

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.

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, and micro array networks in biology; highway transportation networks in civil engineering; topic- or theme-author-publication-citation networks in library science; and wireless telecommunication networks among commanders, soldiers and supply lines in a battle field. In such information networks, each node or link in a network contains valuable, multidimensional information, such as textual contents, geographic information, traffic fellow, and other properties. Moreover, such networks could be highly dynamic, evolving, and inter-dependent.

Traditional data mining algorithms such as classification, market basket analysis, and cluster analysis commonly attempt to find patterns in a dataset containing independent, identically distributed (IID) samples. One can think of this process as learning a model for the node attributes of a homogeneous graph while ignoring the links between the nodes. A key emerging challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets [GD05]. The domains often consist of a variety of object types; the objects can be linked in a variety of ways. 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. 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 [YHYY06], for user-guided clustering across multiple relations [YHY05], for effective link-based clustering [JW02, YHY06], for distinguishing different objects with identical names [YHY07a], and for solving the veracity problem, i.e., finding reliable facts among multiple conflicting web information providers [YHY07b]. 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 [PBMW98] and HITS [Kle99], many variations have been developed to rank one type [CDG+98, Hav02, RD02] or multiple types of objects in the graph [JW02, SQCF05]. 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 [CDI98]. 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) [ACG02, DHM05], natural language processing (co-reference resolution, object consolidation) [PMM+02, BG06], personal information management, and other fields. Recently, there has been significant interest in the use of links for improved entity resolution. The central idea is to consider, in addition to the attributes of the references to be resolved, the other references to which these are linked. These links may be, for example, co-author links between author references in bibliographic data, hierarchical links between spatial references in geo-spatial data, or co-occurrence links between name references in natural language documents. Besides utilizing links in data mining, we may wish to predict the existence of links based on attributes of the objects and other observed links in some problems. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events [OHS05], such as email, telephone calls and co-authorship; and predicting semantic relationships such as \advisor-of" based on web page links and content [CDF+00]. Most often, some links are observed, and one is attempting to predict unobserved links [GFKT01], or there is a temporal aspect.

Another important direction in information network analysis is to treat information networks as graphs and further develop graph mining methods [CH07]. Recent progress on graph mining and its associated structural pattern-based classification and clustering, graph and graph containment indexing, and similarity search will playan important role in information network analysis. An area of data mining that is related to link mining is the work on subgraph discovery. This work attempts to find interesting or commonly occurring subgraphs in a set of graphs. Discovery of these patterns may be the sole purpose of the systems, or the discovered patterns may be used for graph classification, whose goal is to categorize an entire graph as a positive or negative instance of a concept. One line of work attempts to find frequent subgraphs [KK01, YH07], and some other lines of work are on efficient subgraph generation and compression-based heuristic search [WM03, CH07]. 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. Generative models for a range of graph and dependency types have been studied extensively in the social network analysis community [CSW05]. In recent years, significant attention has focused on studying the structural properties of networks [AC05], such as the World Wide Web, online social networks, communication networks, citation networks, and biological networks. Across these various networks, general patterns such as power law degree distributions, small graph diameters, and community structure are observed. These observations have motivated the search for general principles governing such networks [Cha05]. 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 the power of network analysis. 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. Furthermore, in many applications, the facts to be analyzed are dynamic and it is important to develop incremental link mining algorithms. Besides mining knowledge from links, objects and networks, we may wish to construct an information network based on both ontological and unstructured information.

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, sub-sequences, 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.

Frequent pattern mining has been a focused theme in data mining research for over a decade [HCXY07]. 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.

The most focused and extensively studied topic in frequent pattern mining is perhaps scalable mining methods. 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. Recently, studies have proceeded to scalable methods for mining colossal patterns [ZYH+07] 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, and methods for pattern compression and extraction of high-quality top-k patterns [XCYH06]. Much research is still needed to substantially reduce the size of derived pattern sets, mine such patterns directly and efficiently, and enhance the quality of retained patterns.

Moreover, frequent pattern mining could help in other data mining tasks and many such pattern-based mining methods have been developed. Frequent patterns have been used for effective classification by association rule mining (such as [LHM98]), top-k rule generation for long patterns (such as [CTTX05]), and discriminative frequent pattern-based classification [WK05]. Recent studies show that better classification models could be constructed using discriminative frequent patterns and such patterns could be mined efficiently and directly from data [CYHH07, CYHY08]. Frequent patterns have also been used for clustering of high-dimensional biological data [WWYY02]. Therefore, frequent patterns can play an essential role in these major data mining tasks and the benefits should be exploited in depth.

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 like \why this pattern is frequent?" [MXC+07]. 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 [HCXY07, CH07]. Applications often raise new research issues and bring deep insight on the strength and weakness of an existing solution. Much work is needed to explore new

applications of frequent pattern mining, for example, bioinformatics and software engineering.

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.

STREAM DATA MINING

Stream data refers to the data that follows 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 [BBD+02, Agg06].

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 [DH01], load shedding [TcZ+03] and sketching techniques [Mut03], synopsis data structures [GKMS01], stream cubing [CDH+02], and clustering [AHWY03]. Progress has been made on efficient methods for mining frequent patterns in data streams [MM02], multidimensional analysis of stream data (such as construction of stream cubes) [CDH+02], stream data classification [AHWY04], stream clustering [AHWY03], stream outlier analysis, rare event detection [GFHY07], 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.

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 [GHPY02] 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 [MM02] and track the most frequent k items in the continuously arriving data [CM03].

Initial studies on stream clustering concentrated on extending K-means and K-median algorithms to stream environment [GMM+03]. 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. Later, [AHWY03] proposes to divide the clustering process into online microclustering process, which stores summarized statistics about the data streams, and the o²ine one, which performs macro-clustering on the summarized data according to a number of user preferences such as the time frame and the number of clusters. Projected clustering can also be perform for high dimensional data streams [AHWY04].

The focus of stream classification of data streams is first on how to efficiently update the classification model when data continuously flow in. VFDT [DH00] is a representative method in this field where a incremental decision tree is built based on Hoeflding trees. Later, the concept drift problem in data stream classification has been recognized, which refers to the unknown changes of the distribution underlying data streams. Many algorithms have been developed to prevent deterioration in prediction accuracy of the model [HSD01, KM05], by carefully selecting training examples that represent the true concept [Fan04] or combining multiple models to reduce variance in prediction [WFYH03, GFH07]. For skewed distribution of stream data, it is recommended to explore biased selective sampling and robust ensemble methods in model construction [GFHY07].

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.

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, pre-processed 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.

BIBLIOGRAPHY

- [AC05] 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.
- [ACG02] 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.
- [AEEK99] 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.
- [Agg06] C. C. Aggarwal. Data Streams: Models and Algorithms. Kluwer Academic, 2006.
- [AHWY03] 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.
- [AHWY04] 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.
- [All02] James Allan. Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell, MA, USA, 2002.
- [BBD+02] 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.
- [Ber03] M. W. Berry. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2003.
- [BG06] 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.
- [BHLY04] 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.
- [CCLR05] 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.
- [CDF+00] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. Artificial Intelligence, 118:69{113, 2000.
- [CDG+98] S. Chakrabarti, B. E. Dom, D. Gibson, J. M. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In Proc. 7th Int. World Wide Web Conf. (WWW'98), pages 65{74, Brisbane, Australia, 1998.
- [CDH+02] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time series data streams. In Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02), pages 323{334, Hong Kong, China, Aug. 2002.
- [CDH+06] 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.
- [CDI98] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext classification using hyper-links. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data

(SIGMOD'98), pages 307{318, Seattle, WA, June 1998.

- [CH07] D. J. Cook and L. B. Holder. Mining Graph Data. John Wiley & Sons, 2007.
- [Cha01] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In Proc. 2001 Int. World Wide Web Conf. (WWW'01), pages 211{220, Hong Kong, China, May 2001.
- [Cha02] S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002.
- [Cha05] Deepayan Chakrabarti. Tools for large graph mining. PhD thesis, Pittsburgh, PA, USA, 2005. Chair- Christos Faloutsos.
- [CM03] Graham Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. In PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 296{306, New York, NY, USA, 2003. ACM.
- [CSW05] P. J. Carrington, J. Scott, and S. Wasserman. Models and methods in social network analysis. Cambridge University Press, 2005.
- [CTTX05] G. Cong, K.-Lee Tan, A. K. H. Tung, and X. Xu. Mining top-k covering rule groups for gene expression data. In Proc. 2005 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'05), pages 670{681, Baltimore, MD, June 2005.
- [CYHH07] 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.
- [CYHY08] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In Proc. 2008 Int. Conf. Data Engineering (ICDE'08), Cancun, Mexico, April 2008.
- [DH00] P. Domingos and G. Hulten. Mining high-speed data streams. In Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00), pages 71{80, Boston, MA, Aug. 2000.
- [DH01] Pedro Domingos and Geoff Hulten. A general method for scaling up machine learning algorithms and its application to clustering. In ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, pages 106{113, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
- [DHM05] Xin Dong, Alon Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. In SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pages 85{96, New York, NY, USA, 2005. ACM.
- [DM01] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. Mach. Learn., 42(1-2):143{175, 2001.
- [dOL03] M.C. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. IEEE Transactions on Visualization and Computer Graphics, 9(3):378{394, 2003.
- [ER04] Gunes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artificial Intelligence Research, 22:457{479, 2004.
- [Fan04] Wei Fan. Systematic data selection to mine concept-drifting data streams. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 128{137, New York, NY, USA, 2004. ACM.
- [FGW01] U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, 2001.
- [FL05] Federico Michele Facca and Pier Luca Lanzi. Mining interesting knowledge from weblogs: a survey. Data Knowl. Eng., 53(3):225{241, 2005.
- [FLGC02] G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. IEEE Computer, 35:66{71, 2002.
- [GD05] L. Getoor and C. P. Diehl. Link mining: a survey. SIGKDD Explorations, 7:3 { 12, 2005.
- [GFH07] J. Gao, W. Fan, and J. Han. On appropriate assumptions to mine data streams: Analysis and practice. In Proc.

2007 Int. Conf. Data Mining (ICDM'07), Omaha, NE, Oct. 2007.

- [GFHY07] 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.