Rise of Data Mining: Current and Future Application Areas

Dharminder Kumar
[1]Professor and Dean, Faculty of Engineering & Technology,
Guru Jambheshwar University of Science & Technology, Hisar

Deepak Bhardwaj
Research Scholar, Department of Computer Science & Engineering
Guru Jambheshwar University of Science & Technology, Hisar

Abstract

Knowledge has played a significant role on human activities since his development. Data mining is the process of knowledge discovery where knowledge is gained by analyzing the data store in very large repositories, which are analyzed from various perspectives and the result is summarized it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become a very important and guaranteed branch of engineering affecting human life in various spheres directly or indirectly. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern recognition and Computation capabilities have given present day's data mining functionality a new height. Data mining have various applications and these applications have enriched the various fields of human life including business, education, medical, scientific etc. Objective of this paper is to discuss various improvements and breakthroughs in the field of data mining from past to the present and also to explores the future trends.

Keywords— Current and Future of Data Mining, Data Mining, Data Mining Trends, Heterogeneous Data, KDD

I. INTRODUCTION

THE advent of information technology have affected various aspects of human life, may it be in the form of modernization of banking, land records, libraries, or data regarding population. This advent in various fields of human life has led to the very large volumes of data stored in various formats like documents, records, images, sound recordings, videos, scientific data, and many new data formats. An important new trend in information technology is to identify meaningful data collected in information systems [18]. The fact lies in that data is growing at a very rapid rate, but most of data has once been stored and have never been used. This data collected from different sources if processed properly, can provide immense hidden knowledge, which can be used further for development. As this knowledge is captured, it can serve as a key to gaining competitive advantage over competitors in industry [18]. So, there is an eminent need for developing proper mechanisms of processing these large volumes of data and extracting useful knowledge from large repositories for better decision making. Data Mining (as called as Knowledge discovery in databases (KDD), aims at the discovery of useful information from large collections of data [1] but large scale automated search and interpretation of discovered regularities belongs to KDD, but are typically not considered as part of data mining. KDD is concerned with knowledge discovery process applied to databases. KDD refers to overall process of discovering useful knowledge from data, while data mining refers to application of algorithms for extracting patterns from data [18]. The core functionalities of data mining includes applying various methods and algorithms in order to preprocess, classify, cluster and associate the data in order to discover useful patterns of stored data [2]. Data mining is best described as the union of historical and recent developments in statistics, AI, machine learning and Database technologies. These techniques are then used together to study data and find previously-hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which they could not otherwise find [19]. Data mining can be seen as the confluence of multiple fields including statistics, machine learning, databases, pattern discovery and visualization etc. [17]. The various application areas of data mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive advantage, Intelligence, Retail, Finance, Banking, computer, Network, Security, Monitoring, Surveillance, Teaching Support, Climate modeling, Astronomy, and Behavioral Ecology etc. Hence, the objective of this paper is to reviews various trends of data mining and its relative areas from past to present and explores the future areas of it. This paper is organized as follows section 2 presents ground roots of data mining section 3 presents current trends in data mining section 4 presents future trends of data mining and finally conclusion follows.

II. ROOTS OF DATA MINING

Roots of Data Mining can be traced back along three lines[19]. *Final Stage*

A. Statistics

The most important lines is statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

B. Artificial Intelligence & Machine Learning

Data mining's second longest family line is artificial intelligence and machine learning. AI is built upon heuristics as opposed to statistics, and attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. Machine Learning could be considered as an evolution of AI, because it blends AI heuristics with advanced statistical methods. It let computer programs learn about the data they study and then apply learned knowledge to data.

C. Databases

Third family is Databases. Huge amount of data needs to be stored in a repository, and that too needs to be managed. So, comes in light the databases. Earlier data was managed in records and fields, then in various models like hierarchical, network etc. Relational model served the needs of data storage for long while. Other advanced system that emerged are object relational databases. But in data mining, volume of data is too high, so we need specialized servers for it. We call the term as Data Warehousing. Data warehousing also supports OLAP operations to be applied on it, to support decision making [20].

D. Other Technologies

Apart from these, data mining inculcates various other areas, e.g. pattern discovery, visualization, business intelligence etc. The table summarizes the evolution data mining on the grounds of development in databases.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Fig. 1. Evolution of Data Mining, Source [21]

III. CURRENT TRENDS AND APPLICATIONS

Data mining is formally defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [2]. The field of data mining has been growing rapidly due to its broad applicability, achievements and scientific progress, understanding. A number of data mining applications have been successfully implemented in various domains like fraud detection, retail, health care, finance, telecommunication, and risk analysis...etc. are few to name.

The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining; the various challenges include different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. [22]. Advancements in data mining with various integrations and implications of methods and techniques have shaped the present data mining applications to handle the various challenges, the current trends of data mining applications are:

A. Fight against Terrorism [23]

After 9-11 attacks, many countries imposed new laws against fighting terrorism. These laws allow intelligence agencies to effectively fight against terrorist organizations. USA launched Total Information Awareness program with the goal of creating a huge database of that consolidate all the information on population. Similar projects were also launched in European countries and rest of the world. This program faced several problems,

a. The heterogeneity of database, the target database had to deal with text, audio, image and multimedia data.

b. Second problem was scalability of algorithms. The execution time increases as size of data (which is huge). For example, 230 cameras were placed in London, to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per second, which poses heavy loads on both hardware and

software.(Transport for London 2004).

B. Bio-informatics and Cure for Diseases

The second most important application trend, deals with mining and interpretation of biological sequences and structures. Data mining tools are rapidly being used in finding genes regarding cure of diseases like Cancer and AIDS [16].

C. Web and Semantic Web

Web is the hottest trend now, but it is unstructured. Data mining is helping web to be organized, which is called semantic web. The underlying technology is Resource Description Framework (RDF) which is used to describe resources. FOAF is also a supporting technology, heavily used in Facebook and Orkut for tagging. But still there are issues like combining all RDF statements and dealing with erroneous RDF statements. Data mining technologies are serving a lot to make the web, a semantic web.

D. Business Trends

Today's business environment is more dynamic, so businesses must be able to react quicker, must be more profitable, and offer high quality services that ever before. Here, data mining serves as a fundamental technology in enabling customer's transactions more accurately, faster and meaningfully. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends [17]. Almost all of the current business data mining applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong Business Intelligence (BI) system.

IV. DATA MINING – THE NEXT WAVES

Data mining is a promising area of engineering and it does have wide applicability. It can be applied in various domains. 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 [24], [25]. Further Han in research challenges in data mining in science and engineering [26] presents major research challenges in the area of science and engineering.

A. Data Mining in Security and Privacy Preserving

Security and privacy are not very new concepts in data mining, but there is too much that can be done in this area with data mining. [50] gives a thorough analysis of impact of social networks and group dynamics. Specifying the need to understand cognitive networks, he also models knowledge network using the Enron E-mail corpus. Recording of electronic communication like email logs, and web logs have captured human process. Analysis of this can present an opportunity to understand sociological and psychological process. [27] provides various types of privacy breach and present an analysis using k-candidate anonymity [28], [29], k-degree anonymity [30] and k-neighborhood anonymity[31]. Various solutions are emerging like privacy preserving link analysis [32] which needs consideration in future. Secure Multiparty Computation (SMC) [33], [34] can be used where multiple parties, each having private input, want to communicate.

B. Challenges in Mining Financial Data

There are many motivating factors for the study of this area. Biggest is profit[35]. everyone wants profit may it be investor, speculator or operator in trading. He presents models of assets prices, and presents the modeling of relative changes of stock prices. Eraker [36] discuss the issues in modeling stochastic volatility better. [37] present a global solution for Distributed Recommendations in an adaptive decentralized network.

C. Detecting Eco-System Disturbances

This is another promising area. It comprises of many areas such as remote sensing, earth-science, biosphere, oceans and predicts the ecosystem. [38] tries to explain what are the problems in the area and what is the importance. There are also issues in mining the earth science like high dimensionality because long time series data are common in data mining. Study of this area is important due to radical changes in ecosystem has led to floods, drought, ice-storms, hurricanes, tsunami and other disasters [39]. Land Cover Change detection is also one of the areas. In a press release by NASA [40] shows the history of natural disasters.

D. Distributed Data Mining

Conventional data mining is thought to be as containing a large repository, and then mine knowledge. But there is an eminent need for mining knowledge from distributed resources. Typical algorithms which are available to us are based on assumption that the data is memory resident, which makes them unable to cope with the increasing complexity of distributed algorithms [41]. Similar issues also rise while mining data in sensor network, and grid data mining. We need distribution classification algorithms. A technique called partition tree construction approach [42] can be used for parallel decision tree construction. We also need distributed algorithms for association analysis. Distributed ARM algorithms needs to be developed as the sequential algorithms like Apriori, DIC, DHP and FP Growth [43], [44], [45], [46], [47] do not scale well in distributed environment. In his research paper the author presents a Distributed Apriori algorithm [48]. The FMGFI algorithm [49] presents a distributed FP Growth algorithm

V. CONCLUSION

In this paper we briefly reviewed the various data mining trends and applications from its inception to the future. This review puts focus on the hot and promising areas of data mining. Though very few areas are named here in this paper, yet they are those which are commonly forgotten. This paper provides a new perspective of a researcher regarding applications of data mining in social welfare.

REFERENCES

[1] Heikki, Mannila, "Data mining: machine learning, statistics, and databases", *Statistics and Scientific Data Management*, pp. 2-9. 1996.
[2] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. *From Data Mining To Knowledge Discovery in Databases*, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6. MIT 1996.
[3] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age", *in IEEE Intelligent System*s, vol. 14, issue 6, Nov 1999. Doi. 10.1109/5254.809566
[4] Salmin, Sultana et al., "Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture", *International Journal of Multimedia and Ubiquitous Engineering* Vol. 4, No. 4, October, 2009
[5] Hsu J., "Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century", in *The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002)*, ISSN: 1542-7382. Available Online: http://colton.byuh.edu/isecon/2002/224b/Hsu.pdf
[6] Shonali Krishnaswamy, "Towards Situationawareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application", *Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005*, pages-16, 17. Available at : http://www.csse.monash.edu.au/~mgaber/CameraReadyI
[7] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., "Multimedia mining", *WSEAS Transactions on Systems*, No 3, s. 3263-3268, 2005
[8] Abdulvahit, Torun. , Ebnem, Düzgün, "Using spatial data mining techniques to reveal vulnerability of people and places due to oil transportation and accidents: A case study of Istanbul strait", *ISPRS Technical Commission II Symposium*, Vienna. Addison Wesley, 1st edition. 2006
[9] T. M. Mitchell, "Generalization as Search", in *Artificial Intelligence* vol. 18 no. 2, pp.203-226. 1982
[10] R. Michalski., I. Mozetic., J. Hong., and N. Lavrac, "The AQ15 Inductive Leaning System: An Overview and Experiments", *Reports of Machine Leaning and Inference Laboratory*, MLI-86-6, George Maseon University. 1986
[11] J. R. Quinlan. *C4.5: Programs for Machine Learning*, San Francisco: Morgan Kaufmann Publishers, 1993
[12] Z. K. Baker and V. K.Prasanna. "Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs" *IEEE International Parallel and Distributed Processing Symposium* (IPDPS '05), 2005.
[13] Jing He, "Advances in Data Mining: History and Future", *Third international Symposium on Information Technology Application*, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204
[14] Ali Meligy, "A Grid-Based Distributed SVM Data Mining Algorithm", *European Journal of Scientific Research* Vol.27 No.3. Pp.313-321 © Euro Journals Publishing, Inc. Available at :http://www.eurojournals.com/ejsr.htm
[15] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey", *IEEE Trans. Neural Networks*, vol. 13, pp. 3 - 14., 2006
[16] Mark, J., Embrechts, "Introduction to Scientific Data Mining: Direct Kernel Methods & Applications", *Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing*, Wiley , New York, pp. 317-365, 2005
[17] Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.
[18] Walter Alberto Aldana, "Data Mining Industry, Emerging Trends and New Opportunities", Master's Thesis, Massachusetts Institute of Technology, 2000.
[19] Data Mining Software http://www.dataminingsoftware.com
[20] M.S. Chen, J. Han, and P.S. Yu. "Data mining: An overview from database perspective", *IEEE transactions on Knowledge and Data Eng*., 8(6):866-883, December 1999
[21] Pilot Software, "An Introduction to Data Mining", Whitepaper. Pilot Software. 1998.
[22] Venkatadari M., Dr. Lokanataha C. Reddy, "A Review on Data Mining From Past to Future", *International Journal of Computer Applications*, pp. 19-22, vol. 15, No. 7, Feb 2011.
[23] Huysmans, Baesens, Martens, Denys and Vanthienen, "New Trends in Data Mining", *Tijdschrift voor Economie en Management*, vol. L, 4, 2005.
[24] Kargupta, Han, Yu, Motwani, Vipin Kumar, "Next Generation of Data Mining", Chapman & Hall /CRC Data Mining and Knowledge Discovery Series, Taylor and Francis Group LLC, 2008.
[25] J. Han and M. Kamber. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann, San Francisco, CA, 2006.
[26] J. Han, J. Gao, "Research Challenges for Data Mining in Science and Engineering", in Next Generation of Data Mining, Taylor and Francics Group LLC 2008.
[27] Liu, K. Das, T. Grandison & H. Kargupta, "Privacy Preserving Data Analysis on Graph and Social Networks", in Next Generation of Data Mining, Taylor and Francis Group LLC, 2008
[28] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. Technical Report, University of Massachusetts, Amherst,MA, 2007.
[29] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMODSIGART "Symposium on Principles of Database Systems (PODS'98)*, p. 188, Seattle, WA, 1998.
[30] K. Liu and E. Terzi, "Towards identity anonymization on graphs", *In Proceedings of ACM SIGMOD*, pp. 93–106, Vancouver, Canada, June 2008.
[31] B. Zhou and J. Pei. "Preserving privacy in social networks against neighborhood attacks" In *Proceedings of the 24th International Conference on Data Engineering (ICDE'08)*, pp. 506–515, Cancun, Mexico, April 2008.
[32] Y. Duan, J. Wang, M. Kam, and J. Canny, "Privacy preserving link analysis on dynamic weighted graph" *Computational and Mathematical Organization Theory*, 11:141–159, 2005.
[33] A. C. Yao " How to generate and exchange secrets", In *Proceedings of 27th IEEE Symposium on Foundations of Computer Science*, pp. 162–167, Toronto,Canada, 1986.
[34] O. Goldreich. The Foundations of Cryptography, Vol. 2, Chapter 7. Cambridge University Press, Cambridge, UK, 2004.
[35] James E. Gentle, "Challenges in Financial Data Mining", In Next Generation of Data Mining, Taylor and Francis Group, LLC 2008.
[36] Eraker, B., 2004, "Do stock prices and volatility jump? Reconciling evidence from spot and option prices", *The Journal of Finance* 59, 1367–1403.
[37] Olfa Nasraoui and Maha Soliman, "Market-Based Profile Infrastructure: Giving Back to the User", Next Generation of Data Mining, Taylor and Francis, 2008.
[38] Shayam Boriah, Vipin Kumar, Michael Steinbach, Pang-Ning Tan, Christopher Potter, and Steven Klooster, "Detecting Ecosystem Disturbances and Land Cover Change using Data Mining", In NGDM, 2008
[39] C. Potter, P.-N. Tan, V. Kumar, C. Kucharik, S. Klooster, V. Genovese, W. Cohen, and S. Healey. "Recent history of large-scale ecosystem disturbances in North America derived from the AVHRR satellite record". *Ecosystems*, 8(7):808–824, 2005.
[40] Press Release: "Data Mining Reveals a New History of Natural Disasters",NASA. http://www.nasa.gov/centers/ames/news/releases/2003/03 51AR.html.
[41] Chris Clifton, Wei Jiang, M. Murugesan, and M.E. Nergiz, "Is Privacy Still and Issue for Data Mining", In NGDM, Taylor and Francis, 2008.
[42] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, pp. 23–34, Atlanta, GA, April 2006.
[43] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In *Proceedings of the 20th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 247–255, Santa Barbara, CA, May 21–23, 2001. ACM.
[44] W. Kim, R. Agrawal, C. Faloutsos, U. Fayyad, J. Han, G. Piatetsky-Shaprio, D. Pregibon, and R. Uthurasamy. "Data mining" is not against civil liberties. Open Letter from the Directors of the Executive Committee of ACMSIGKDD, July 28, 2003.
[45] A. Kobsa. "Technical solutions for Privacy-enhanced personalization", *Communications of the ACM*, 50: 24–33, August 2007.

[46] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities, I (281): 31–50, October 24, 1995.
[47] L. Sweeney. "k-anonymity: A model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5): 557–570, 2002.
[48] C. Clifton. "Using sample size to limit exposure to data mining", *Journal of Computer Security*, 8(4): 281–307, November 2000.
[49] R. Feingold, J. Corzine, R. Wyden, and B. Nelson. Data Mining Moratorium Act of 2003. U.S. Senate Bill (proposed), January 16, 2003.
[50] N. Pathak, S. Mane, J. Srivastava, N. Contractor, S. Poole & D. Williams, "Analysis of Social Networks and Group Dynamics from Electronic Communication", In Next Generation of Data Mining, Taylor and Francics, 2008.