Research Paper Engineering

Current and Future Applications of Data Mining

Dhathrika Sai Ganesh ECM Department, Sreenidhi Institute of Science & Technology, JNTUH, Hyderabad, India

D. Charishma ECM Department, Sreenidhi Institute of Science & Technology, JNTUH, Hyderabad, India

ABSTRACT

Knowledge has played a major role on individual activities since his development. Data mining (DM) is the process of knowledge discovery where knowledge is gained by analyzing the data store in huge repositories, which are analyzed from various perceptions and the outcome is summarized it into useful information. Due to the significance of extracting knowledge from the large data repositories, DM has become a very vital and certified branch of engineering affecting individual life in various fields directly or indirectly. Advancements in Statistics, Machine Learning (ML), Artificial Intelligence (AI), Pattern recognition(PR) and Computation capabilities have given current day's DM functionality a new stature. DM have various applications and these applications have developed the various fields of individual life including business, education, medical, scientific etc. The main objective of this paper is to discuss different improvements and advancements in the field of DM from past to the present and also to discovers the future trends.

KEYWORDS Data Mining, CRM, Heterogeneous Data, KDD

INTRODUCTION

The initiation of information technology (IT) have affected various aspects of individual life, may it be in the form of transformation of banking, land records or data regarding population. This initiation in various fields of individual life has led to the very huge volumes of data stored in different formats like documents, records, images, sound recordings, videos, scientific data, and lots of new data formats. An important new trend in IT is to identify significant data collected in ISs. The fact lies in that data is rising at a very rapid rate, but most of data has once been stored and have never been used. This data collected from diverse sources if processed properly, can provide huge hidden knowledge, which can be used further for development. As this knowledge is captured, it can serve as a key to gaining cutthroat advantage over competitors in industry. So, there is an important need for developing proper mechanisms of processing these huge volumes of data and extracting useful knowledge from huge repositories for better decision making. DM aims at the innovation of useful information from large collections of data but large scale automated search and understanding of discovered regularities belongs to KDD, but is typically not considered as part of DM **[1]**. KDD is concerned with knowledge discovery process applied to databases. KDD refers to overall process of discovering useful knowledge from data, while DM refers to application of algorithms for extracting patterns from data.

The core functionalities of DM includes applying various techniques and algorithms in order to preprocess, classify, cluster and associate the data in order to discover useful patterns of stored data **[2, 3]**. DM is best described as the union of historical and recent developments in statistics, AI, machine learning and Database technologies **[8]**. 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. Data mining can be seen as the confluence of multiple fields including statistics, ML, pattern discovery (PD) and visualization etc. The various application areas of DM are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications **[4]**, Manufacturing, Retail, Finance, Banking, Security, Monitoring, Surveillance, Climate modeling and Behavioral Ecology etc. Hence, the aim of this paper is to reviews diverse trends of DM **[5]** and its virtual areas from past to present and explores the future areas of it.

ORIGIN OF DATA MINING

Statistics

The most significant lines are statistics. Without statistics, there would be no DM, as statistics are the foundation of most technologies on which DM is built **[6]**. Statistics hold concepts such as regression analysis (RA), standard distribution (SD), standard deviation (SDv), standard variance (SV), discriminant analysis (DA), cluster analysis (CA), and, all of which are used to study data and their relationships. These are the very building blocks with which more advanced SAs are emphasized. Definitely, within the heart of today's DM tools and methods, classical SAs play an important role.

Artificial Intelligence

DM second greatest family line is AI and ML. AI is built upon heuristics as opposed to statistics, and attempts to apply human thought like processing to statistical difficulties. Because this approach needs enormous computer processing power, it was not realistic until the early 1985s. AI found a few applications at the very high end scientific markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. ML could be considered as an evolution of AI, because it combines AI heuristics with advanced statistical techniques **[8]**. It allows computer programs study about the data they study and then apply learned knowledge to data.

Databases

The third family is Databases (DBs). Enormous amount of data needs to be stored in a repository, and that too needs to be managed. So, comes in light the DBs. Earlier data was managed in records and fields, then in various replicas like hierarchical, network etc. Relational model served the needs of data storage space for long while. Other advanced system that emerged is object relational databases (ORDBs). But in DM, volume of data is too elevated, so the specialized servers are needed for it and it is called as Data Warehousing (DW). Data warehousing also supports OLAP operations to be applied on it, to carry decision making.

New Technologies
Apart from these, DM inculcates various other fields, e.g. PD, visualization, BI etc. The table 1 summarizes the evolution DM on the grounds of development in DBs.

Table 1: Evolution of Data Mining

Evolutionary Step	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	Computers, Disks, Tapes	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	RDBMS, Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, Dynamic data delivery at record level
Data warehousing & Decision Support (1990s)	On-Line analytical processing (OLAP), Multidimensional databases, Data warehouses	Pilot, Cognos, Micro strategy	Retrospective, Dynamic data delivery at multiple level
Data Mining (Emerging Today)	Advanced algorithms, Multiprocessor computers, Massive databases	Pilot, IBM, SGI	Prospective, proactive information delivery

CURRENT TRENDS AND APPLICATIONS
DM is formally defined as the non-trivial procedure of identifying valid, new, potentially useful, and ultimately understandable patterns in data. The field of DM has been growing rapidly due to its broad applicability, achievements and scientific progress, understanding **[7]**. A number of DM applications have been successfully implemented in various domains like fraud detection, retail, health care, finance, telecommunication, and risk analysis...etc.

The ever increasing difficulties in various areas and improvements in technology have posed new challenges to DM; a variety of challenges consist of different data formats, data from dissimilar locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. Advancements in DM with various integrations and inferences of methods and techniques have shaped the current DM applications to handle the various challenges, the current trends of DM applications are:

Battle against violence
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 enormous database of that merge all the information on population. Similar pilot projects were also launched in European countries and rest of the world. This agenda faced several problems,

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

The second difficulty was scalability of algorithms. The execution time increases as size of data.

Bio-informatics and Cure for Diseases
The second most significant application trend, deals with mining and interpretation of biological progression and structures. DM tools are rapidly being used in finding genetic materials regarding cure of diseases like Cancer and AIDS.

Web and Semantic Web
Web is the hottest development now, but it is shapeless. DM is helping web to be organized, which is called semantic web (SW). The primary technology is Resource Description Framework (RDF) which is used to describe resources. FOAF is also a supporting knowledge, heavily used in Facebook and Orkut for tagging. But still there are problems like combining all RDF statements and dealing with invalid RDF statements. DM technologies are serving a lot to make the web as semantic web **[10]**.

Business Trendz
Today's business background is more dynamic, so businesses must be able to respond quicker, must be more profitable, and offer high quality services that ever before. Here, DM serves as a primary technology in enabling customer's transactions more accurately, faster and meaningfully. DM techniques of classification, regression, and CAs are used for in present business trends. Almost all of the present business DM applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong BI system.

DATA MINING – THE NEXT SWING
DM is a capable area of engineering and it does have wide applicability. It can be applied in diverse domains DM, as the convergence of multiple intertwined disciplines, including statistics, ML, PR, DBs, information retrieval (IR), World Wide Web (WWW), visualization, and many application fields, has made great development in the earlier period. The research challenges in DM and engineering presents major research challenges in the field of science and engineering.

Data Mining in Security and Privacy
Security and privacy are not very new concepts in DM, but there is too much that can be done in this area with DM. It gives a careful analysis of impact of social networks and group dynamics. Specifying the need to recognize cognitive networks, he also models knowledge network using the Enron E-mail body. Recording of electronic communication like E-Mail logs, and web logs have captured human process. Analysis of this can present a chance to understand sociological and psychological procedure. It provides different types of privacy breach and presents an analysis using k-candidate ambiguity, k-degree ambiguity and k-neighborhood ambiguity. A variety of solutions are emerging like privacy preserving link analysis which needs consideration in future. Secure Multiparty Computation (SMC) can be used where multiple parties, each having confidential input, want to communicate.

Detecting Eco-System Annoyances
This is another promising area. It comprises of many fields such as remote sensing (RS), earth-science (ES), biosphere, oceans and predicts the system. They try to explain what are the difficulties in the area are and what the importance is. There are also issues in mining the earth science like high dimensionality because long time series data are common in DM. Study of this area is important due to radical changes in ecosystem has led to overflows, drought, ice-storms, tsunami and other disasters.

Distributed Data Mining
Conventional DM is thought to be as containing a huge repository, and then mine knowledge. But there is an important need for mining knowledge from distributed resources. Typical algorithms which are available to us are based on hypothesis that the data is memory resident, which makes them unable to cope with the increasing difficulty of distributed algorithms. Similar issues also rise while mining data in sensor network, and grid DM. The distribution classification algorithms are needed. A technique called partition tree construction approach can be used for parallel decision tree (DT) construction **[9]**. The distributed algorithms for association analysis are also needed. Distributed ARM algorithms needs to be developed as the sequential algorithms like Apriori, DIC, DHP and FP Growth do not scale well in distributed surroundings. In his research paper the author presents a Distributed Apriori algorithm. The FMGFI algorithm presents a distributed FP Growth algorithm

CONCLUSION
In this paper, the various DM trends and applications from its beginning to the future are briefly reviewed. This review puts

focus on the hot and promising areas of DM. Though very few areas are named here in this paper, yet they are those which are commonly forgotten. This paper provides a new perception of a researcher regarding applications of DM in social welfare.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Management of Sreenidhi Institute of Science & Technology, Hyderabad for their constant encouragement and co-operation.

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 Systems*, 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 Situation awareness 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/CameraReady
7. 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
8. J. R. Quinlan. *C4.5: Programs for Machine Learning,* San Francisco: Morgan Kaufmann Publishers, 1993
9. 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.
10. 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