Research Paper

Computer Science

Data Mining Techniques and Trends – A Review

Shital H. Bhojani Assistant Professor-Anand Agriculture University-Anand - Gujarat

Dr. Nirav Bhatt Assistant Professor – RK University, Rajkot - Gujarat

ABSTRACT

Everyday Terabytes of data are generated in many organizations. So it's difficult to predict for the world. Because of data are increasing day by day, we requires a need for new tools and techniques to support humans in automatically and intelligently analyzing large data repositories to obtain useful information. These growing needs gives a vison for a new area of research field called Data Mining (DM) or Knowledge Discovery in Databases (KDD).DM aims to extract implicit, previously unknown and potentially useful information from data by digging intelligently in large data repositories. In another way we can say that DM techniques are needed /used to extract unknown predictive information from large mass of data. Now a days data mining enhanced the various fields of human life including business, education, agriculture, medical, scientific etc., using Artificial Intelligence, Statistics, Computation capabilities, Pattern Recognition and Machine Learning, data visualization techniques. So we can say that DM has become an essential component in various fields of human life.This paper discusses and describes DM and major DM techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithm, and visualization.

KEYWORDS : Knowledge Discovery in Databases (KDD), Data Mining, Trends, Association, Classification, Clustering, Prediction, pattern recognition.

INTRODUCTION

Data and information plays a very vital role on human activities. The knowledge discovery process is as old as Homo sapiens. Until some time ago this process was solely based on the 'natural personal' computer provided by Mother Nature. Fortunately, in recent decades the problem has begun to be solved based on the development of the Data mining technology, aided by the huge computational power of the 'artificial' computers.DM is an active research area and research is ongoing to bring statistical analysis and artificial intelligence (AI) techniques together to address the issues. DM is the search for valuable information in large volumes of data (Weiss and Indurkhya, 1998). It is the process of nontrivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques (Technology Forecast, 1997; Piatetsky-Shapiro and Frawley, 1991).

DATA MINING

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge. In general, there are three main steps in DM: preparing the data, reducing the data and, finally, looking for valuable information. The specific approaches, however, differ from companies to companies and researchers to researchers. Fayyad et al. (1996) proposed the following steps:

1. Retrieving the data from a large database.
2. Selecting the relevant subset to work with.
3. Deciding on the appropriate sampling system, cleaning the data and dealing with missing fields and records.
4. Applying the appropriate transformations, dimensionality reduction, and projections.
5. Fitting models to the preprocessed data.

DATA MINING TECHNIQUES

There are several major data mining techniques have been developing and using in data mining projects including association, classification, clustering, prediction, pattern recognition.

Association Rules

Association rule is one of the well-known data mining technique. It implies certain association relationships among a set of objects in a database. Association rule mining is generally performed in generation of frequent Item sets. Nowadays Retailers are using association technique to research customer's buying habits to provide best services to customers and increase sales.

Classification

Classification is the practices of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification predicts discrete, unordered labels. Mainly classification technique is used to categorize data item into several predefined set of classes or groups. Any prediction can be thought of as classification or estimation. Different types of Classification method are:

Classification by decision tree induction

Bayesian Classification

It's a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG)

Neural Networks

An artificial neural network (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and/or functional aspects of biological neural networks.

Support Vector Machines (SVM)

Support vector machine (SVM) is a set of related supervised learning methods used for classification and regression. SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

Classification Based on Associations

Prediction

Regression technique can be adapted for predication in data mining. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In prediction records are classified according to some predicted future behavior or projected future value. In data mining independent variables are attributes already known and response variables are what we want to predict. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Types of regression methods

- Linear Regression

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Clustering

Clustering makes valuable cluster of objects. Clustering can be said as identification of similar subgroups or clusters of objects. By using clustering technique, we find some kinds of similarities in one cluster and label it with a meaningful name. The records are grouped together on the basis of self-similarity. Many clustering algorithms have been developed and are categorized from several aspects such as Partitioning Methods, Hierarchical Agglomerative, (divisive) methods; Density based methods, Grid-based methods, and Model-based methods.

General Types of Clusters are Well-separated clusters, Center-based clusters, Contiguous clusters, Density-based clusters

Pattern Recognition:

There are many applications of machine learning in pattern recognition. One is optical character recognition, which is recognizing character codes from their images. This is an example where there are multiple classes, as many as there are characters we would like to recognize. Especially interesting is the case when the characters are handwritten. People have different handwriting styles; characters may be written small or large, slanted, with a pen or pencil, and there are many possible images corresponding to the same character. In the case of face recognition, the input is an image, the classes are people to be recognized, and the learning program should learn to associate the face images to identities. This problem is more difficult than optical character recognition because there are more classes, input image is larger, and a face is three-dimensional and differences in pose and lighting cause Significant changes in the image.

DIFFERENT TRENDS OF DATA MINING

The diversity of data, data mining tasks, and data mining approaches poses many challenging research issues in data mining. The development of efficient and effective data mining methods and systems and techniques to solve large application problems are important tasks for DM researchers and data mining system and application developers. The era of data mining applications was conceived in the year1980 primarily by research-driven tools. There are a number of data mining trends is in terms of technologies and methodologies which are currently being developed and researched. The early day's data mining trends are as under.

DATA TRENDS

Previously, flat files, traditional and relational databases were used to store the data which uses tabular representation. Data mining algorithms work best for numerical data collected from a single data base, and various data mining techniques have evolved. Later on, with the union of Statistics and Machine Learning techniques, various algorithms evolved to mine the non-numerical data and relational databases.

COMPUTING TRENDS

Computing is the need of world. The field of data mining has been greatly influenced by the fourth generation programming languages and related computing techniques. In the previous era of data mining maximum of the algorithms working only on statistical techniques. Later on they evolved with different computing techniques like artificial intelligence, machine learning and pattern reorganization.

CURRENT TRENDS

Now a days DM has been become very popular due to its tremendous success in terms of broad-ranging application achievements and scientific progress. The ever increasing complexities in various fields and enhancements in technology have posed new challenges to world of data mining; to handle the various challenges, some of the current trends are as under:

DISTRIBUTED/COLLECTIVE DATA MINING (CDM)

In distributed data mining data are located in different places, in different physical locations. The man goal of CDM is to effectively mine distributed data which are located in heterogeneous locations. CDM provides a better approach to vertically partitioned datasets, using the notion of orthonormal basis functions, and computes the basis coefficients to generate the global model of the data (Kargupta et. al., 2000).

HYPERTEXT AND HYPERMEDIA DATA MINING

Hypertext and hypermedia data mining can be characterized as mining data which includes text, hyperlinks, text markups, and various other forms of hypermedia information. Some of the important data mining techniques used for hypertext and hypermedia data mining include classification (supervised learning), clustering (unsupervised learning), semi-structured learning, and social network analysis.

CONSTRAINT- BASED DATA MINING

This form of data mining incorporates the use of constraints which guides the process.Frequently this is combined with the benefits of multidimensional mining to add greater power to the process (Han, Lakshamanan, and Ng, 1999). There are several categories of constraints which can be used, each of which has its own characteristics and purpose.

UBIQUITOUS DATA MINING

The advent of laptops, palmtops, cell phones, and wearable computers is making ubiquitous access to large mass of data possible. The Ubiquitous computing environments are subsequently giving rise Ubiquitous Data Mining (UDM). UDM is the process of analysis of data for extracting useful knowledge from the data of ubiquitous computing. Accessing and examining data from a ubiquitous computing device may offer many challenges.

MULTIMEDIA DATA MINING

Multimedia Data Mining is the mining and study of several types of data, including images, audio, video and animation. Some of the DM techniques that are applied on multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, also association rule mining, clustering methods. It's new filed of research, but holds much potential for the future.

SPATIAL DATA MINING

The spatial data includes astronomical data, natural resources data, satellite data and space craft data. Some of the data mining techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubes, spatial OLAP, and spatial clustering methods. Mostly these data are of image-oriented, and can represent a great deal of information if properly analyzed and mined (Miller and Han, 2001).

TIME SERIES DATA MINING

It focuses on the goal of identifying movements or components which exist within the data sets of stock prices, currency exchange rates, the volume of product sales, biomedical measurements, weather data, etc (trend analysis). These can include long-term or trend movements, seasonal variations, cyclical variations, and random movements (Han and Kamber, 2001). Some of the rule induction algorithms Version Space, AQ15, C4.5 rules are presently employed in Time series data mining applications.

BUSINESS TRENDS

Early data mining applications focused mainly on helping businesses gain a competitive edge. The exploration of data mining for businesses continues to expand as e-commerce and e- marketing have become mainstream elements of the retail industry. Today's business/ industry must be more cost-effective, very faster and offer high value services that ever before.

Due to customer's expectations and constraints, data mining becomes a fundamental technology in supporting customer's transactions more accurately. Most probably classification and prediction Techniques are used for supporting business decisions and progressed to Decision Support Systems (DSS) and very recently it has grown to Business Intelligence (BI) systems.

FUTURE TRENDS

Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments. Ever increasing technology and future application areas are always poses new challenges and opportunities for data mining, the typical future trends of data mining includes:

- Standardization of data mining languages
- Data preprocessing
- Complex objects of data
- Computing resources
- Web mining
- Scientific Computing
- Business data

Conclusion

In closing, it would not be overly optimistic to say that, DM will be one of the main viable focuses of the world. Although improvements are continuously been made in the DM field, many issues remain to be resolved and much research has yet to be done. The capability to continually change and provide new thoughtful is the principle benefit of DM, and will be at the core of DM bright and promising future. Having the right information at the right time is essential for making the right decision. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. In the millennium, world will be competing in generating information from large mass of data rather than collecting data.

References

1. Salmin, Sultana et al. 2009. Ubiquitous Secretary: AUbiquitous Computing Application Based on WebServices Architecture , International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4,October, 2009
2. Hsu, J. 2002. Data Mining Trends and Developments:The Key Data Mining Technologies and Applications for the 21st Century, 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
3. Kotsiantis, S., Kanellopoulos, D., Pintelas, P. 2004. Multimedia mining. WSEAS ransactions on Systems,No 3, s. 3263-3268.
4. T. M. Mitchell. 1982. Generalization as Search, Artificial Intelligence, 18(2), 1982, pp.203-226.
5. R. Michalski., I. Mozetic., J. Hong., and N. Lavrac. 1986. The AQ15 Inductive Leaning System: An Overview and Experiments, Reports of Machine Leaning and Inference Laboratory, MLI-86-6, George Maseon University.
6. J. R. Quinlan.1992.Programs for Machine Learning, Morgan Kaufmann.
7. Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.
8. Kargupta, H. et al, "Collective Data Mining," in Advances in Distributed Data Mining, Karhgupta and Chan, editors, MIT Press, 2000.
9. Kargupta, H. and A. Joshi, "Data Mining To Go: Ubiquitous KDD for Mobile and Distributed Environments," Presentation, KDD-2001, San Francisco, August 2001.
10. Data Mining :Concepts, Models and Techniques: Authors: Gorunescu, Florin
11. Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan, Kaufmann Publishers
12. Data mining techniques and applications – A decade review from 2000 to 2011
13. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsia, Department of Management Sciences, Tamkang University, Taiwan, ROC
14. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012 : Data Mining and Its Applications for KnowledgeManagement : A Literature Review from 2007 to 2012. Tipawan Silwattananusarn, Dr. KulthidaTuamsuk Khon Kaen University, Thailand
15. REVIEW OF LITERATURE ON DATA MINING Mrs. Tejaswini Abhijit Hilage1 & R. V. Kulkarni 2 IJRRAS 10 (1) ● January 2012 www.arpapress.com/Volumes/Vol10Issue1/IJRRAS_10_1_14.pdf
16. A Review Paper on Various Data Mining Techniques Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat Amit Khaparde Volume 4, Issue 4, April 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
17. A review of data mining techniques: Sang Jun Lee - University of Nebraska-Lincoln, Lincoln, Nebraska, USA, Keng Siau - University of Nebraska-Lincoln, Lincoln, Nebraska, USA
18. International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011 : A Review on Data mining from Past to the Future Venkatadri.M Dr. Lokanatha C. Reddy K Suguna et al, Int.J.Computer Technology & Applications,Vol 6 (4),583-585
19. LITERATURE REVIEW ON DATA MINING TECHNIQUES K.Suguna Asst.Professor, Dr.K.Nandhini Professor
20. Piatetsky-Shapiro, Gregory. 2000. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems.