Survey on Current Trends and Techniques of Data Mining Research

Syed romat Ali [a],Sohail anwar [b] ,Ali zaib khan[c] , Munib Ahmmad [d]

romatali512@gmail.com

1. ABSTRACT

The paper surveys various aspects of data mining research. Data mining is helpful in acquiring knowledge from large domains of databases, data warehouses and data marts. Various and current fields of data mining were also discussed. Various open source tools as well as data mining issues and challenges are also addressed. Data mining is an important and developed research field and is also used by biologists for statisticians and computer scientists.

Keywords: *Data mining, knowledge discovery in databases, areas and tools in data mining, challenges of data mining.*

2. INTRODUCTION

Data mining is extracting information and knowledge from huge amounts of data. Data mining is an essential step in knowledge discovery from databases. There are number of databases, data warts, data warehouses worldwide. If the data is not analyzed to detect interesting patterns, then the data will become a data grave. Data miners look for pearls in a sea of data. A data mining system can generate a lot of patterns. A small part of the pattern is usually interesting. Interesting here means usable, valid and novel. Furthermore, it is almost impossible to extract interesting hidden patterns in the sea of data without the help of data mining tools.

Data mining consists of seven stages. They are: data cleaning, data integration, data selection, data transformation, data mining and knowledge current-action and pattern development. Database technology evolved from primitive file processing to the development of data mining tools and applications. Data can be collected from various applications including science and engineering, management, business houses, government administration and environmental control. Interesting data patterns can be mined from spatial, time-related, textual, biological, multimedia, web, and legacy databases. Data mining facilitates management in decision making. A data mining job includes concept discovery,

association, classification, prediction, clustering, trend analysis, deviation analysis and similar analytics analysis. Data mining in large databases presents various requirements and challenges for researchers and developers. A multidimensional data model is used for the design of data warehouses and data carts. The basic data of such a model is the cube. A data cube has a large set of numbers of facts and dimensions. Dimensions are the entities on which an organization keeps records. By nature, they are hierarchical.

3. DIFFERENT AREAS OF DATA MINING

3.1 Web Mining

As there is an enormous amount of data and information available in the World Wide Web, data miners have a fertile field for web mining. Web mining is a data mining technique for extracting information from web documents and services. The contents of the web are very dynamic. It is growing at a rapid pace, and information is constantly being updated. Web mining can be divided into the following sub-types.

1. Resource Search: Search the desired document for the web.
2. Information Selection and Preprocessing: Selecting and Preprocessing Information Retrieved from the Web.
3. Generalization: Searching for common patterns from multiple sites simultaneously with the individual.
4. Analysis: Discovered patterns are interpreted for meaningful knowledge. Web mining may be divided into Web Structure, Web Contents, and Web Access Designs.

3.2 Text Mining

The term text mining or KDT (Knowledge Discovery in Text) was first proposed in 1996 by Feldman and Dagen. Unstructured text can be mined using information retrieval, text classification, or NLP techniques as a preprofitting step. Text mining involves many applications such as text categorization, clustering, patterns and sequential patterns in lessons, computational linguistics, and search for associations.

3.3 Spatial Data Mining

Spatial data mining relates to location-related data. The explosion of data related to geographic for the rapid development of IT, the demand for GIS to develop databases for digital mapping, remote sensing, spatial analysis and modeling. Spatial data description, classification, association, clustering, trending and external analysis are the main components of spatial data mining.

3.4 Multimedia Data Mining

Multimedia data mining discovers interesting patterns from multimedia-related databases that manage a large collection of multimedia objects. Multimedia objects include text, text markup and linkage with audio, video,

images, sequence data, and hypertext data. Multimedia data research focuses on content-based retrieval, similarity search, association and classification and prediction analysis.

3.5 Time Series Data Mining

The time series database changes its values and events with respect to time. Some examples of time series data are stock market data, business transaction data, dynamic production data, medical treatment data, webpage access sequences and soon. Time series research includes issues related to similarity search, trend analysis, mining sequential and periodic patterns in time-related data.

3.6 Biological Data Mining

There is a large storage of clinical and biological data from DNA microarray data, genomic sequences, protein interactions as well as sequences, electronic health records, disease pathways, biomedical images and the list goes on. In a clinical context, biologists are trying to find the biological processes that cause a disease. There are some issues related to these high-dimensional biological data. These cases involve noise and incomplete data, integrating various sources of data and processing computer intensive tasks. Biologists as well as clinical scientists used a variety of data mining tools to discover interesting and meaningful observations from a large number of heterogeneous data from different biological domains.

3.7 Educational Data Mining

Educational data mining (EDM) is an emerging research field that deals with unique types of data that come from educational settings, and uses those methods to better understand students. Educational data mining focuses on developing new tools and algorithms for data pattern discovery. EDM develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning. New computer-supported interactive learning methods and tools have opened up opportunities to gather and analyze student data, discover patterns and trends in those data, and make new discoveries and hypotheses about learning to students. Data collected from online learning systems can be collected by a large number of students and may include many variables that data mining algorithms can search for model building. Different student models are used to predict students' future learning behavior. Computational models are used based on student domains and pedagogy.

3.8 Ubiquitous Data Mining (UDM)

Data miners have a new challenge in the form of ubiquitous access using wearable computers, palmtops, cell phones, laptops. Advanced analysis is required to extract hidden information from these devices. In the world of UDM, communication,

computing, security, etc. are some of the factors. One of the objectives of UDM is to find interesting patterns while reducing the additional cost of computing due to the above pattern. Implementing data mining functions such as classification, clustering, associations, etc. is difficult for ubiquitous tools. Small display area, data management in mobile has some challenges in this regard. The major issues are advanced algorithms for mobile and distributed computing, data management issues, data representation techniques, integration of these devices with database applications, UDM architecture, software agents, agent interaction and applications of UDM.

3.9 Constraint-Based Data Mining

Constraint-based data mining is one of the developing areas where data miners use constraint for better data mining. One of the applications of constraint-based data mining is the Online Analytical Mining Architecture (OALM) developed by [6] and is designed for multi-dimensional as well as construction based mining based on data bases and data warehouses. has been done. Typically, data mining techniques lack user control. One form of data mining is one where human participation occurs in the form of barriers. There are various types of constraints with their own characteristics and purpose. They are knowledge type, data, dimension / level, interestingness, rule constraints.

4. DATA MINING TOOLS

The following are popular data mining tools:

4.1 Rapid Miner:

The tool is written in the Java programming language, and provides advanced level analysis through its template-based framework. Users rarely have to do any coding. Rapid Minor is capable of handling a variety of tasks such as statistical modeling, predictive analysis and visualization as part of data mining tasks. Rapid Minor provides learning plans, models and algorithms from WEKA and R scripts that make it more powerful. It is distributed under the open source AGPL open source license and can be downloaded from Source Forge. It is one of the best business analytics software.

4.2 WEKA

Weka was originally developed in a non-Java version for analysis of agricultural data. Later, the Java version was developed, and it became a powerful tool for various data mining applications such as predictive modeling and data analysis. The software is free under the GNU General Public License, which is a major advantage over Rapid Minor. As it is free under the GNU General Public License which is a major advantage compared to its counterparts like Rapid Mine. It can be customized by users. Most data mining jobs are supported by Weka. They are classification, clustering, regression, feature extraction, visualization, etc. Its graphical user interface makes it a sophisticated tool

for the data mining process. Therefore, Weka has become one of the most powerful open source data mining software.

4.3 R Programming Project R,

It is a GNU project, written in C, FORTRAN and R Language. R language is used to write many modules of software. Programming software is free, and is also used for statistical computing and graphics. Data miners used R to develop statistical packages and analyze data. In recent years the popularity of R had increased due to its ease of use and use. R provides different statistical techniques including linear and non-linear modeling; Data mining process ie classification, clustering, time series analysis and others.

4.4 Orange

Orange is a Python-based, powerful and open source tool aimed at data extraction users for data mining. It has powerful visual programming and associated Python scripting. It can be used by adding add-ons for machine learning as well as bioinformatics and text mining. It is full of features for data analytics. Orange has special add-ons such as Bio Orange for bioinformatics.

4.5 KNIME

KNIME is capable of performing three main functions in data preprocessing. They are extraction, transformation and loading. Data processing is done by allowing the assembly of nodes. It is an integration platform with robust data analytics and reporting. KNIME used the modular data pipelining concept for machine learning and data mining. It is used for business intelligence as well as financial data mining. KNIME is easily expandable and plug-ins can be added for specific jobs. It is also written in open source Java and is based on Eclipse. The main version includes various data integration modules. Its research area includes not only pharmaceutical research but also business data, financial intelligence and CRM customer data.

4.6 NLTK

When it comes to language processing tasks, NLTK is one of the major players. NLTK is used for machine learning, data mining, sentiment analysis, and data scraping. It is also used extensively for language processing. Because it is written in Python, anyone can build applications on top of it, optimizing it for smaller tasks. NLTK played a major role as learning tool, study tool, prototype and can be used as a platform for high quality research.

5. LITERATURE REVIEW

There are a lot of data mining studies around the world.

The authors used the distribution features of the text classification that took into account the compactness and position of the first appearance of the word. Previous researchers used a bag of representations of researchers' words and gave a word with values and were concerned about whether

the word appeared in the document or the frequency of the word. In their research work, the authors discovered other types of values that express the distribution of a word in a document. Distribution features are used by the IDF style equation and features from different categories are combined using joint learning techniques. The authors experimentally proved that distribution attributes are useful for text classification. In contrast to traditional methods, using these attributes with additional features significantly improves classification performance. The performance of delivery features is enhanced in the case of longer documents and when the writing style is contingent.

The authors designed the web service recommendation system. When designing web service recommendation systems, the focused research problem was to avoid recommending inappropriate or poor services to users. The system should help users choose the right service from a large number of available web services. A widely recommended metric in this regard is the reputation of web services. The feedback rating service by users is used to provide a reputation score. Malicious and subjective user feedback often leads to biases that affect the reputation measurement of web services. In his research work, he proposed a novel system for the same. Cumulative sum control charts and Pearson correlation coefficients were used to find malicious user feedback ratings. The system performed better using Bloom filtering and the proposed Malicious Response Rating Prevention Plan. Extensive experiments were conducted using 1.5 million web service call records. Experimental results showed that the ratio of success of web service recommendations can be increased and the deviation of system reputation measurement can be reduced. The researchers proposed a novel intelligent system that would be able to automatically detect road accidents, inform them using a network of vehicles, and estimate the severity of an accident based on data mining tools and knowledge interventions. Various variables such as vehicle speed, type of vehicles involved, impact speed, and airbag condition, etc. are used to measure the severity of the accident. A prototype based on off-the-shelf equipment was developed and validated at the Applus + IDIADA Automotive Research Corporation facilities, which showed that the system would alert and deploy emergency services after an accident May not reduce the time required. Three classification algorithms were used such as decision trees, support methods

6. DATA MINING TECHNIQUES

Many data mining techniques are used in data mining tasks. Association, classification, clustering, prediction, sequential pattern mining, etc. are data mining techniques.

6.1 Classification

Classification discovers rules that partition data into certain groups. The input for

classification is the training set. Classroom labels of the training set are already known. Classification provides labels to non-recorded records based on a model that derives knowledge from the training dataset. Such classification is known as supervised learning because class labels are known. There are several classification models. Some of the common classification models are decision trees, neural networks, genetic algorithms, support of vector machines, Bayesian classifiers. Applications include credit risk analysis, fraud detection, banking and medical applications, etc.

6.2 Clustering

Clustering is a method of grouping data so that the data within a cluster has a high uniformity and spread of data to other clusters. Clustering algorithms can be used for organizing data, classifying data for model building and data compression, external detection, etc. Many clustering algorithms were developed and segmentation methods, hierarchical methods, density based and grid methods are classified as Datasets can be numeric or hierarchical. K-Means, Hierarchical, DBSCN, Optics, STIN are some famous data cluster algorithms.

6.3 Association Rule Mining

Association rule mining is a well-researched method for discovering interesting relationships between variables in large databases. In the union rule, the expression is of the form X => Y, where the X and Y items are set. The main objective is to search for all rules whose support and trust is greater than or equal to the minimum support or trust in the database. Support means how often X and Y occurs together as a percentage of total transactions. Confidence refers to how dependent a particular object is on another. There is no significance to the pattern with low confidence and support. Users can extract useful and interesting information from patterns with intermediate values of trust and support.

6.4 Neural Networks

Neural networks are new computing paradigms that are inspired by biological nervous systems, such as the brain, to process information. It involves developing mathematical structures with the ability to learn. Neural networks have the ability to extract meaningful and useful patterns and trends from complex data. This particularly applies to real-world problems in the case of industry. Since neural networks are good at identifying patterns or trends, they may be applicable to prediction or front casting needs. This system is made up of highly interconnected processing elements (neurons) working together to solve a particular problem. For example, artificial neural network (ANN) learns. ANN is configured for classification, pattern recognition, etc. for a specific application through a learning process. It can also be used for three-dimensional object recognition, hand-written word recognition,

facial recognition, etc. Neural networks have the drawback of not interpreting derived results. Another problem is that it suffers from long learning times. As the data grows, the situation gets worse for that problem.

6.5 Support Vector Machines

Support vector machines (SVMs) belong to a new class of machine learning algorithms and are based on statistical learning theory. The main concept is to set the data non-linearly in a high dimensional feature space and use a linear differentiator for the classification of the data. It is basically used for regression, classification and decision tree construction. SVMs select the plane that maximizes the margin separating the two classes. The margin is defined as the distance between the hyper plane to the nearest point of A, as well as the distance from the hyper plane to the nearest point of B, where A and B are two linearly different sets. SVM has been used in many applications including face recognition, handwritten character and digit recognition, speech recognition, image and information retrieval.

6.6 Genetic Algorithms

Genetic algorithm is a new paradigm, inspired by Darwin's theory of evolution. The population of the individual is initially randomly constructed with a possible solution to a problem. The crossover is then performed by mixing pairs of individuals to produce the next generation of offspring. A mutation process is used to randomly modify the genetic structure of some members of the new generation. The algorithm searches for a solution in sequential generation. When an optimal solution is found or a certain amount of time elapses, the process is terminated. Genetic algorithms are widely used in problems where optimization is required.

7. ACKNOWLEDGEMENT

The author expressed his gratefulness to Prof. Shahan Yamin Siddiqui, Vice-Chancellor, Dr. Sajid Mehmood Shahzad, Minhaj University Lahore. The authors also acknowledged Prof. Muhammad Saleem Akhtar, Head of Department of Computer Science for his valuable suggestions.

REFERENCES

1. Adam Baba, Gauz Pasha, Shaik Altaf Ahmed, s. Naseera Tabassum, "Introduction to Neural Networks Design Architecture", International Journal of Scientific and Engineering Research Volume 4, Issue 2, Fubury 2013, ISS 2229-5518.

2. Priests Arun, Data Mining Techniques, University Press, 2013.

3. Christos n. Moridis and Anastasios a. Econoids "Mood Recognition During Online Self Assessment Tests" on IEEE Transaction Technologies, VOL. 2, no. 1, JANUARY March 2009.

4. Eric Hondesh-Chan Lu, Wang-Chien Li, Member, IEEE, and Vincent S. Teng, Member, IEEE, "A Framework for Personal Mobile Commerce Pattern Mining and Prediction", IOE Transmissions on KNOWLEDED and Data Engineering, VOL. 26, no. 5, May 2012.

5. H. Karagupta and A. Joshi, "Data Mining to Go: Ubiditus KDD for Mobile and Distributed Environments", KDD-2001, San Francisco, August 2001.

6. J. Han, VS. Lakshmanan and R. T. N. G., "Constraint-based, multidimensional data mining", Computer (Special issue on data mining), 32 (8): 45-50, 1999.

7. Jiawei Han and Michaline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufman Publishers, 2003.

8. Kasun Vikra Maratha, Student Member, IEEE, Miroslav Kubat, Senior Member, IEEE, and Kamal Premaratne, Senior Member, IEEE, "Preview of Missing Items in Shopping Cart", IEEE Parts on KNOWLEDGE and Data Engineering, IOL. 21, No. 7, July 2009.

9. Li-der Chow, Member, IEEE, Nien-Hwa Lai, Yen-Wen Chen, Member, IEEE, Yao-Jen Chang, Xuan-Yan Yang, Lien-Fu Huang, Wen-Ling Chiang, Hung-Yi Chiyu, and Haw-Yun Shin "IEEETRANSACTIONS on Information Technology" Services for Children's Families with Mobile Social Network Development Services.

10. Luigi Lanceri, Member, IEEE and Nicholas Durand "Internet User Behavior: Access Trace and Application of the Discovery of Communities", IEEE Transactions Compared to Systems, Man, and Siberics - Part A: Systems and Humans, VOL. 37, No. 1, JANUARY 2006.

11. Manuel Fogg, Piedad Garrido, Member, IEEE, Francisco J. Martinez, Member, IEEE, Juan-Carlos Cano, Carlos T. Calafet, and Pietro Manzoni, Member, IEEE, "Systems for Automatic Notification and Severity of Automotive" Accidents ", IEEE Transportation Mobile Computing, VOL. 13, No. 5, May 2014.

12. 12. Maya Nayak and Gyan Ranjan Tripathi: "Pattern Classification Using Neuro Fuzzy and Support Vector Machines (SVM) - A Comparative Study", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, issue 5, may 2013

13. N. Malumbo, "Data Mining: Techniques, Key Challenges and Approaches to Improvement", International Journal of

Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

14. Pasco Konjevoda and Nikola štambuk, "Open-source tools for data mining in the social sciences," Theoretical and methodological approaches to social science and knowledge management, pp. 163–176.

15. Shanggang Wang, Member, IEEE, Jibin Zheng, Member, IEEE, Zhengping Wu, Member, IEEE, Fungchun Yang, Member, IEEE, Michael R. Leeu, Fellow, IEEE, "Reputation Measurement and Malicious Response Rating Prevention in Web Services.