Survey on Current Trends and Techniques of Data Mining Research

Sadiq Hussain

ABSTRACT

The paper surveys different aspects of data mining research. Data mining is helpful in acquiring knowledge from large domains of databases, datawarehouses and data marts. Different and current areas of data mining are also discussed. Issues and challenges of data mining along with different open source tools are addressed as well. Data mining is an important and evolving research area and used by the biologists to statisticians and computer scientists as well.

London Journals Press

London Journal of Research in Computer Science and Technology

Volume 17 | Issue 1 | Compilation 1.0

2 0 U K

Survey on Current Trends and Techniques of Data Mining Research

Sadiq Hussain

_____

I. ABSTRACT

The paper surveys different aspects of data mining research. Data mining is helpful in acquiring knowledge from large domains of databases, data warehouses and data marts. Different and current areas of data mining also discussed. Issues and challenges of data mining along with various open source tools are addressed as well. Data mining is an important and evolving research area and used by the biologists to statisticians and computer scientists as well.

Author: System Administrator, Dibrugarh University, Dibrugarh, India.

II. INTRODUCTION

Data mining is extracting information and knowledge from huge amount of data. Data mining is an essential step in discovering knowledge from databases. There are numbers of databases, data marts, data warehouses all over the world. If the data are not analyzed to find out the interesting patterns, then the data would become data tombs. Data miners seek for the pearl in the sea of data. A data mining system may generate lots of patterns. Typically a small fraction of the patterns are interesting. Here the interesting means useable, valid and novel. Moreover, it is almost impossible to extract the interesting hidden patterns in the sea of data without the help of data mining tools. There are seven steps in data mining. They are data cleaning, data integration, data selection, data transformation, data mining, knowledge presentation and pattern evolution[7]. Database technology had evolved from primitive file processing to the development of data mining tools and applications. The data may be collected from various applications including science and engineering, management, business houses, government administration and environmental control. Interesting data patterns may be mined from spatial, time-related, text, biological, multimedia, web and legacy databases. Data mining facilitate management in decision making. The data mining job includes the discovery of concept descriptions, association, classification, prediction, clustering, trend analysis, deviation analysis and similarity analysis. Data mining in large databases poses various requirements and challenges for the researchers and developers. A multidimensional data model is used for the design of data warehouses and data marts. The core of such model is data cube [7]. Data cube consists of large set of facts and number of dimensions. Dimensions are the entities on which an organization keeps records. By nature, they are hierarchical.

III. DIFFERENT AREAS OF DATA MINING

3.1 Web Mining

As there is huge amount of data and information available in the World Wide Web, the data miners have a fertile area for web mining. Web mining is data mining techniques for extraction of information from web documents and services. The contents of the web are very dynamic. It is growing at a rapid pace, and the information is

continuously updated. Web mining may be divided into the following subtasks [2].

1. *Resource finding:* finding documents intended for the Web.

2. *Information selection and preprocessing:* Selection and preprocessing of the information retrieved from the Web.

3. *Generalization:* To discover the general patterns from the individual as well as multiple sites.

4. *Analysis:* Discovered patterns are interpreted for meaningful knowledge.

Web mining may be divided into Web Structure, Web Contents, and Web Access Patterns.

3.2 Text Mining

The term text mining or KDT (Knowledge Discovery in Text) was first proposed by Feldman and Dagan in 1996 [2]. The unstructured text may be mined using information retrieval, text categorization, or applying NLP techniques as a preprocessing step. Text Mining involves many applications such that text categorization, clustering, finding patterns and sequential patterns in texts, computational linguistics, and association discovery.

3.3 Spatial Data Mining

The spatial data mining deals with data related to location. The explosion of geographically related data for rapid development of IT, digital mapping, remote sensing, GIS demands for developing databases for spatial analysis and modeling. Spatial data description, classification, association, clustering, trend, and outlier analysis are the main components for spatial data mining.

3.4 Multimedia data mining

Multimedia data mining explores the interesting patterns from databases related to multimedia that manages a large collection of multimedia objects. Multimedia objects include audio, video, image, sequence data and hypertext data containing text, text markups, and linkages.

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

3.5 Time series data mining

A time series database changes its values and events with respect to time. Some of the examples of time series data are stock market data, business transaction data, dynamic production data, medical treatment data, web page access sequence and so on. The time series research involves 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 the clinical context, biologists are trying to find the biological processes that are the cause of a disease. There are some issues related to these high-dimensional biological data. These matters include noisy 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 area concerned with the unique types of data that come from educational settings, and using those methods to better understand students. Educational Data Mining focuses on developing new tools and algorithms for discovering data patterns. 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 collect

and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building. Different student models are used for prediction of future learning behavior of the students. Computational models are used based on the student domain and pedagogy.

3.8 Ubiquitous data mining (UDM)

The data miners have a new challenge in the form of the ubiquitous access by using wearable computers, palmtops, cell phones, laptops. To extract hidden information from these devices requires advanced analysis. In the world of UDM, communication, computation, security, etc. are some of the factors. The one of the objectives of the UDM is to extract interesting patterns while minimizing the additional cost of the computing due to the above-cited factors. To implement data mining tasks like classification, clustering, associations, etc. are difficult for ubiquitous devices. Small display areas, data management in mobile are some of the challenges in this regards. The key issues are the advanced algorithm 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 [5].

3.9 Constraint-based data mining

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

IV. DATA MINING TOOLS

The following are the popular data mining open source tools.

4.1 RapidMiner

This tool is written in Java programming language, and it offers analytics of advanced level through its template-based framework. Users hardly have to do any coding. RapidMiner is capable of handling various tasks like statistical modeling, predictive analytics and visualization apart from data mining tasks. Rapid- Miner provides learning schemes, models and algorithms from WEKA and R scripts that make it more powerful. This open source is distributed under the AGPL open source license and it can be downloaded from SourceForge. It is one of the best business analytics software. All the data mining tasks are bundled in one single suite [http://rapid-i.com/content/view/181/ 190/].

4.2 WEKA

Weka was originally developed in a non-Java version for analyzing agricultural data. Later, the Java version was developed, and it became a powerful tool for different data mining applications like predictive modeling and data analysis. This software is free under the GNU General Public License, which is a big advantage compared to RapidMiner. As it is free under the GNU General Public License which is a big advantage of it as compared to its counterparts like RapidMiner. It can be customized by the users. Most of the data mining jobs are supported by Weka. They are classification, clustering, regression, feature extraction, visualization, etc. Its graphical user interface makes it a better-sophisticated tool for data mining process. So, Weka has become one of the most powerful open source data mining software. [http://en.

wikipedia.org/wiki/Weka_(machine_learning)] [http://www.cs.waikato.ac.nz/ml/ weka/].

4.3 R-Programming

Project R, which is a GNU project, is written in C, FORTRAN and R Language. R language is used for writing lots of modules of the software itself. R programming software is free, and it is also used for statistical computing and graphics. Data miners used R for developing statistical packages and analyzing the data. In recent years the popularity of R had increased because of its ease of use and extensibility. R provides different statistical techniques that include linear and nonlinear modeling; data mining processes i.e. classification, clustering, time series analysis and others. [http://www.r-project.org/][14].

4.4 Orange

Orange, a Python-based, powerful and open source tool for data mining users for the purpose of knowledge extraction. It has powerful visual programming and Python scripting attached to it. It can be used for machine learning as well as bioinformatics and text mining by adding add-ons. It's packed with features for data analytics. Orange has specialized add-ons like Bioorange for bio-informatics [http://orange.biolab.si/ features/].

4.5 KNIME

KNIME is capable of performing three main tasks in data preprocessing. They are extraction, transformation, and loading. The data processing is done by allowing the assembly of nodes. It is an integration platform with strong data analytics and reporting. KNIME used 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 extendible and can be added a plug-in for specific jobs. This open source is also written in Java and based on Eclipse. The core version consists of various data integration modules. Its research area not only includes pharmaceutical research but also business data, financial intelligence and

CRM customer data. [https://en.wikipedia.org/ wiki/KNIME].

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 extensively used for language processing. Because it's written in Python, one can build applications on top of it, customizing it for small tasks. NLTK played a major role as a teaching tool, study tool, prototyping and can be used as a platform for high-quality research. [https://en.wikipedia.org/ wiki/Natural_Language_Toolkit]

V. LITERATURE REVIEW

There are lots of data mining studies around the globe.

Students Mood recognition [3] was proposed by Christos N. Moridis et. al. for online self-assessment test. Exponential logic and formulas were used in this regards. The inputs were student's previous answers and slide bar status. The exponential logic variables were a total number of questions for the online self-assessment test, student's goal, and slide bar value. Appropriate feedbacks are recorded based on current status of moods of the students. Student's manual selection of their mood using slide bar without any automation is the limitation of the system.

A novel weakly supervised cyber criminal network mining method [18] was proposed by Raymond Y.K. Lau et. al. The technique was based on relationships both explicit and implicit among the cyber criminals. The messages posted by these criminals on the social media were the basis of this method. The algorithm used in this context was context-sensitive Gibbs sampling algorithm. The algorithm mined both transactional and collaborative semantics to find the relationship among such criminals. The model used was a probabilistic generative model for extracting multi-word expressions. Two types of cyber

criminal relationships were established in unlabeled messages. The approach used here is concept level for the implicit semantics associated with the text.

Shenghua Bao et. al. [16] proposed for discovering and connecting with social emotions based on the online documents with emotions to help the users to select related documents by their emotional preferences. This is a problem of document categorization. For such social affective text mining, a joint emotion-topic model was proposed by introducing an additional layer for such kind of emotion modeling into Latent Dirichlet Allocation (LDA). Associate emotions with specific emotional context were used instead of a single term. The authors developed an approximate inference model by using Gibbs Sampling Algorithm. The model categorized text based on different emotions such as touch, surprise, and empathy, etc. by using social affective text as input.

Luigi Lancieri et. al. [10] proposed a classification method for Internet users based on their behavior at net to offer enhanced services. For this purpose, IP Address, timestamp, keywords from proxy cache, URL, categorized user behaviour were collected. Two different kinds of categorization algorithms were used. One is called "hard clustering" for partition and another is "soft clustering" for finding overlapping clusters to group users. Hierarchical agglomerative clustering (HAC) was used for hard clustering.

Li-Der Chou et. al. [9] proposed the use of social media with the help of mobile devices to create social network group for the children with developmental disabilities (CDD). Families with CDD, university, hospital and foundation came hand to hand to share significant information based on online social network related to childcare of such children. The users can access the application with the help of PDA, personal computer or mobile devices by installing the application on such devices.

In [17], the authors used distributional features of text categorization that took into account the compactness and the position of the first appearance of the word. Previous researchers had used 'bag of words' representation and assigned a word with values and concerned with whether the word appeared in the document or not or the frequency of the word. The authors in their research work explored other types of values which express distribution of word in a document. The distributional features are used by a tf idf style equation and features of different categories are combined using ensemble learning techniques. The authors proved experimentally that distributional features are useful for text categorisation. The categorisation performance improves significantly by using these features with little additional cost in contrast to traditional methods. The distribution features performances are enhanced the case of long documents and when the writing style is casual.

In [15], the authors designed web service recommendation systems. While designing web service recommendation systems, the focused research problem was to avoid recommending unfair or poor services to the users. The system should help users to choose right service from the huge number of available web services. The widely recommended metric in this regards is the reputation of web services. The feedback ratings by the users are used for providing service reputation score. Malicious and subjective user feedback often leads to bias that affects the reputation measurement of web services. In their research work, they proposed a novel system for the same. Cumulative Sum Control Chart and Pearson Correlation Coefficient were used to find malicious user feedback ratings. The system performed better by using Bloom filtering and proposed malicious feedback rating prevention scheme. Extensive experiments were conducted by using 1.5 million web service invocation records. The experimental results showed that success ratio of the web service recommendations may be enhanced and the system might reduce the deviation of reputation measurement.

In [11], the researchers proposed a novel intelligent system which would be able to detect the road accidents automatically, notify them by using vehicular networks and estimate the severity of the accident based on data mining tools and knowledge interference. Various variables such as the vehicle speed, the type of vehicles involved, the impact speed, and the status of the airbag, etc. are used for measuring the severity of the accident. A prototype based on off-the-shelf devices was developed and validated it at the Applus + IDIADA Automotive Research Corporation facilities, showing that this system can reduce the time needed to alert and deploy emergency services notably after an accident takes place. Three classification algorithms were used such as Decision Trees, Support Vector Machines, and Bayesian networks and were compared for best results. It was found that Bayesian model for classification is the best-suited model.

In [4], the authors proposed a novel system called Mobile Commerce Explorer (MCE). The system was for mining as well as prediction of mobile users' movements. It can also be used for purchasing transactions under the context of mobile commerce. The framework (MCE) contains three major components - 1)Similarity Inference Model (SIM) for measuring the similarities among stores and items, which are two basic mobile commerce entities considered in the paper; 2) Personal Mobile Commerce Pattern Mine (PMCP-Mine) algorithm for efficient discovery of mobile users' Personal Mobile Commerce Patterns (PMCPs) and 3) Mobile Commerce Behavior Predictor (MCBP) for prediction of possible mobile user behaviors. The study predicted mobile users' commerce behaviors to recommend stores and items previously not known to a user.

In [8], the researchers proposed a technique for the prediction of what else the customer likely to buy based on partial information about the contents of a shopping cart. The data structure used in this context was itemset trees (ITtrees), they obtained all the rules whose antecedents contain at least one item that is missing from the shopping cart in a computationally efficient manner. The classical Bayesian decision theory and a new algorithm based on Dempster-Shafer (DS) theory of evidence combination were combined for finding out rules based uncertainty processing technique. The proposed algorithm enhanced the performance. As the input, the algorithm takes an incoming item set and returns a graph based on association rules entailed by the incoming item set. The proposed algorithm used depth-first search technique and also updated the rule graph.

VI. DATA MINING TECHNIQUES

Several 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 finds rules that partition data into some groups. The input for the classification is the training set. The training set's class labels are already known. Classification assigns class labels to unlabelled records based on a model that acquires knowledge from the training datasets. Such classification is known as supervised learning as the class labels are known. There are several classification models. Some of the common classification models are decision trees, neural networks, genetic algorithms, support vector machines, Bayesian classifiers. The application includes credit risk analysis, fraud detection, banking and medical application, etc. [2].

6.2 Clustering

Clustering is a method of grouping data so that data within the cluster have high similarity and dissimilar to data in other groups. Clustering algorithms may be used for organizing data, categorize data for model construction and data compression, outlier detection, etc. Many clustering algorithms were developed and are categorized as partitioning methods, hierarchical methods, density based and grid based methods.

The datasets may be numerical or categorical. K-Means, hierarchical, DBSCAN, OPTICS, STING are some of the well-known data clustering algorithms [13].

6.3 Association Rule Mining

Association rule mining is a well-researched method for discovering interesting relations between variables in large databases. In association rule, the expression is of the form X=>Y, where X and Y are set of items [2]. The main objective is to discover all the rules that have support and confidence greater than or equal to minimum support or confidence in a database. Support means that how often X and Y occurs together as a percentage of total transactions. Confidence means that how much a particular item is dependent on another. There is no significance for the patterns with low confidence and support. The users can extract useful and interesting information from the patterns with intermediate values of confidence and support. The association rule mining algorithms include Apriori, AprioriTid, Apriori hybrid and Tertius algorithms [13].

6.4 Neural Networks

Neural networks are new computing paradigm that is inspired by the biological nervous system, such as the brain, to process information [13]. It involves developing mathematical structures with ability to learn [2]. The Neural networks have the ability to extract meaningful and useful patterns and trends from the complex data. It is applicable to real world problems especially in case of industry. As the neural networks are good at identifying patterns or trends, they may be applicable for prediction or forecasting needs. The system is composed of highly interconnected processing elements (neurons) working together to solve a specific problem. Artificial neural network (ANN) learns by example [15]. ANN is configured for specific application as classification, pattern recognition etc. through a learning process. It may also be used for three-dimensional object recognition, hand-written word recognition, face recognition, etc. Neural networks have the drawback of not explaining the derived results. Another problem is that it suffers from long learning times. As the data grows, the situation becomes worse for that problem.

6.5 Support Vector Machines

Support vector machines (SVM) belong to a new class of machine learning algorithms and are based on statistical learning theory [2]. The main concept is to non-linearly map the data set into a high dimensional feature space and use a linear discriminator for classification of data. It is basically used for regression, classification and decision tree construction. SVMs select the plane which maximizes the margin separating the two classes. The margin is defined as the distance between the separating hyperplane to the nearest point of A, plus the distance from the hyperplane to the nearest point in B, where A and B are two linearly separable sets. SVM has been used in many applications including face detection, handwritten character and digits recognition, speech recognition, image and information retrieval [12].

6.6 Genetic Algorithms

Genetic algorithms are a new paradigm in computing inspired by Darwin's theory of evolution [2]. A population of the individual with possible solution to a problem is created initially at random. Then the crossover is done by combining pairs of individuals to produce offspring of next generation. A mutation process is used to modify the genetic structure of some members of new generation randomly. The algorithm searches for a solution in the successive generation. When an optimum solution is found or some fixed time is elapsed, the process comes to an end. Genetic algorithms are widely used in problems where optimization is required.

VII. ACKNOWLEDGEMENT

London Journal of Research in Computer Science and Technology

Dibrugarh University for his inspiring words. The author also acknowledged Prof. Gopal Chandra Hazarika, Professor, Department of Mathematics for his valuable suggestions.

REFERENCES

1. Adam Baba, Gouse Pasha, Shaik Althaf Ahammed, S. Nasira Tabassum, "Introduction to Neural Networks Design Architecture", International Journal of Scientific & Engineering Research Volume 4, Issue 2, Februry 2013, ISSN 2229-5518.

2. Arun K Pujari, Data Mining Techniques, University Press, 2013.

3. Christos N. Moridis and Anastasios A. Economides "Mood Recognition during Online Self-Assessment Tests" IEEE Transactions on Learning Technologies, Vol. 2, NO. 1, January March 2009

4. Eric Hsueh-Chan Lu, Wang-Chien Lee, Member, IEEE, and Vincent S. Tseng, Member, IEEE," A Framework for Personal Mobile Commerce Pattern Mining and Prediction", IEEE Transactions on Knowledge And Data Engineering, Vol. 24, No. 5, May 2012.

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

6. J. Han, V.S. Lakshmanan and R T Ng, "Constraint-based, Multidimensional Data Mining", COMPUTER (Special issue on Data Mining), 32(8): 45-50, 1999.

7. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2003.

8. Kasun Wickramaratna, Student Member, IEEE, Miroslav Kubat, Senior Member, IEEE, and Kamal Premaratne, Senior Member, IEEE," Predicting Missing Items in Shopping Carts", IEEE Transactions on Knowledge And Data Engineering, Vol. 21, No. 7, July 2009.

9. Li-Der Chou, Member, IEEE, Nien-Hwa Lai, Yen-Wen Chen, Member, IEEE, Yao-Jen Chang, Jyun-Yan Yang, Lien-Fu Huang, Wen-Ling Chiang, Hung-Yi Chiu, and Haw-Yun Shin "Mobile Social Network Services for Families With Children With Developmental Disabilities" IEEE Transactions on Information Technology.

10. Luigi Lancieri, Member, IEEE, and Nicolas Durand "Internet User Behavior: Compared Study of the Access Traces and Application to the Discovery of Communities" IEEE Transactions On Systems, Man, and Cybernetics—part A: Systems And Humans, Vol. 36, No. 1, January 2006.

11. Manuel Fogue, Piedad Garrido, Member, IEEE, Francisco J. Martinez, Member, IEEE, Juan-Carlos Cano, Carlos T. Calafate, and Pietro Manzoni, Member, IEEE," A System for Automatic Notification and Severity Estimation of Automotive Accidents", IEEE Transactions on Mobile Computing, Vol. 13, No. 5, May 2014.

12. Maya Nayak and Jnana Ranjan Tripathy: "Pattern Classification Using Neuro Fuzzy and Support Vector Machine (SVM) – A Comparative Study", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013.

13. N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

14. Paško Konjevoda and Nikola Štambuk, "Open-Source Tools for Data Mining in Social Science," Theoretical and Methodological Approaches to Social Sciences and Knowledge Management, pp. 163-176 .

15. Shangguang Wang, Member, IEEE, Zibin Zheng, Member, IEEE, Zhengping Wu, Member, IEEE, Fangchun Yang, Member, IEEE, Michael R. Lyu, Fellow, IEEE, "Reputation Measurement and Malicious Feedback Rating Prevention in Web Service

16. Recommendation Systems", IEEE Transactions on Services Computing, Vol. , No. , March 2014.

17. Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu "Mining Social Emotions from Affective Text "IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 9, September 2012.

18. Xiao-Bing Xue and Zhi-Hua Zhou, Senior Member, IEEE, "Distributional Features for Text Categorization", IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 3, March 2009.

19. Y.K. Raymond, Lau SAR Yunqing Xia, Yunming Ye Shenzhen "A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media", China.

London Journal of Research in Computer Science and Technology

This page is intentionally left blank