International Journal of Engineering Research ISSN:2319-6890 (online),2347-5013(print)
Volume No.8, Issue No.4, pp : 114-118 1 Oct. 2019

Current Areas and Challenges of Data Mining Research

Shaik Mahammad Rafi, ObiliNarendra Reddy, Naveen Kumar B.
Assistant Professors, Department of Computer Science and Engineering,
Annamacharya Institute of Technology and Sciences::Rajampet, AP, India

Abstract: The paper describes 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.

Keywords: data mining, knowledge discovery in databases, areas and tools in data mining, challenges of data mining.
Author: System Administrator, Dibrugarh University, Dibrugarh, India.

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 pearlintheseaofdata.Adataminingsystemmay 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 decisionmaking. 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.

Different Areas of Data Mining

Web Mining

As there is huge amount of data and information available in the World Wide Web, the dataminers 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 theWeb.
3. Generalization: To discover the general patterns from the individual as well as multiple sites.
4. Analysis: Discovered patterns are interpreted for meaningfulknowledge.

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

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 associationdiscovery.

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 datamining.

ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

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, and classification and prediction analysis.

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, webpage access sequence and soon. The time series research involves issues related to similarity search, trend analysis, mining sequential and periodic patterns in time-related data.

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.

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 discoveringdatapatterns.EDMdevelopsmethods 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.

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].

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.

Rapid Miner: 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. Rapid Miner is capable of handling various tasks like statistical modeling, predictive analytics and visualization apartfromdataminingtasks.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 Source Forge. 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/].

ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

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 Rapid Miner. As it is free under the GNU General Public License which is a big advantage of it as compared to its counterparts like Rapid Miner. 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/].

4.1 *R-Programming*

Project R, which is a GNU project, is written in C, FORTRAN and R Language. R language is used forwritinglotsofmodulesofthesoftwareitself.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].

4.2 *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 Bio orange for bio-informatics [http://orange.biolab.si/features/].

4.3 *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].

4.4 *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. LIRERETERATUVIEW

ShenghuaBaoet.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.

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.

ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

VI. DATA MININGTECHNIQUES

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 somegroups.Theinputfortheclassificationisthe 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 RuleMining*

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 mainobjectiveistodiscoveralltherulesthathave 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 *NeuralNetworks*

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 applicableforpredictionorforecastingneeds.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 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 VectorMachines*

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 *GeneticAlgorithms*

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 algorithmsearchesforasolutioninthesuccessive 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 isrequired.

REFERENCES

i. *Adam Baba, Gouse Pasha, Shaik Althaf Ahammed, S. Nasira Tabassum, "Introduction to Neural Networks Design Architecture", International Journal of Scientific & Engineering Research Volume*

ISSN:2319-6890 (online),2347-5013(print)
1 Oct. 2019

4, Issue 2, Februry 2013, ISSN 2229-5518.

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

iii. Christos N. Moridis and Anastasios A. Econo- mides "Mood Recognition during OnlineSelf- Assessment Tests" IEEE Transactions On Learning Technologies, Vol. 2, No. 1, January March2009

iv. 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, May2012.

v. H. Kargupta and A. Joshi, "Data Mining to Go: Ubiquitous KDD for Mobile and Distributed Environments", KDD-2001, San Francisco, August2001.

vi. 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.

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

viii. KasunWickramaratna, Student Member, IEEE ,Miroslav Kubat, Senior Member, IEEE, and Kamal Premaratne, Senior Member, IEEE".

ix. 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" IEEETRANSAC- TIONS ON INFORMATIONTECHNOLOGY.

x. 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, January2006.

Authors Proposed by:

1. Shaik Mahammad Rafi,
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences:Rajampet

2. ObiliNarendra Reddy
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences:Rajampet

3. Naveen Kumar B
Assistant Professor,
Department of C.S.E,
Annamacharya Institute of Technology and Sciences:Rajampet