Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Data Mining Application & Tools

[1]V. S. Jagadheeswaran, [2]V. N. Saranya

[1] Assistant Professor, Department of Information Technology, Dr N.G.P Arts and Science College, Coimbatore, India

[2] M.Phil Research scholar, Department of computer science, Dr N.G.P Arts and Science College, Coimbatore, India

Abstraction: Data mining is the process of analyzing data from different outlooks and summarizing it into useful information. Information, that can be used to increase earnings, cuts costs, or both. Data mining software having a number of analytical tools for analyzing data. It allows users to analyze data from different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process, which helps to find correlations or patterns among dozens of fields in large relational databases.

Keywords: Data, Cluster, Tools, information, Tools.

I. INTRODUCTION

We can use data mining techniques in basic form to tuff task. Each techniques differ slightly in purpose and goal. Intrinsic, data mining helps organizations to analyze copies amount of data in order to detect common patterns or learn new things. Without automation we cannot process all these data. **Here some examples:**

- Cluster detection is a type of pattern recognition that is used to detect patterns within large data classes. It is like arranging a large amount of information into categories using patterns and data analysis.
- Anomaly detection helps to find abnormalities in data. It is used in many areas, such as weather patterning and forensic computing.
- Regression is a technique that helps to predict future outcomes using large sets of existing variables. It is used to predict user engagement, customer retention and even property prices.

DATA, INFORMATION, AND KNOWLEDGE

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating data in different formats and different databases. This includes:

- Operational or transactional data is nothing but It is about, sales, cost, inventory, payroll, and accounting
- Nonoperational data is about industry sales, forecast data, and macro economic data
- Meta data is a data about the data itself, such as logical database design or data dictionary definitions

Information

Data of patterns, associations, or relations can provide information. For example, analysis of retail point of sales transaction, data can yield information on which products are selling in copies amount and when it sold.

Knowledge

Knowledge is about copies of information. It tells about historical patterns and future trends. For example, summarizing information on retail supermarket sales, analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouses

Dramatic results in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is a process of centralized data management and retrieval. Data warehousing and data mining, is a relatively new term the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a place where something, especially a natural resource, is found in significant quantities of all data. Activities of an data is needed to maximize user access and analysis. Tremendous technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data easily. The data analysis software is one of the process, which supports even data mining.

II. TECHNIQUES

This provides a elucidation of some of the most common data mining algorithms use today. Attic Techniques: Statistics, Neighborhoods and Clustering.

Statistics

"Statistics" or statistical techniques are not data mining. Statistics is a long term method which is used before the term data mining. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective will be faced with a conscious choice when solving a "data mining" problem, we can attack it with statistical methods or other data mining techniques. For this reason it is important to have some idea of how statistical techniques work and how they can be applied.

Nearest Neighbor

Clustering and the Nearest Neighbor prediction technique are used in data mining. Most people have an intuition that they understand what is clustering, which means records are grouped or clustered together. Nearest neighbor is a prophecy technique that is quite similar to clustering - its essence that in order to predict what a prophecy value in one record, look for records with similar predictor values in the historical database and use the prediction value from the record that it “nearest” to the unclassified record.

Clustering

Clustering is the method of records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is used to mean in segmentation - which helps marketing people, let to know how useful for coming up with a birds eye view of the business. Two of these clustering systems are the PRIZM™ system from Claritas corporation and Micro Vision™ from Equifax corporation. These companies have grouped the population by demographic information into segments that they believe are useful for direct marketing and sales. To build these groupings they use to gathering informations which includes income, age, occupation, housing and race collect in the US Census. Then they naming memorable “nicknames” to the clusters.

III. DATA MINING - APPLICATIONS

Data Mining is widely used in various areas. There are number of commercial data mining system available, though there are many challenges in this field. Here is the list of areas where data mining is broadly used:

- Medical and pharma
- Insurance and Healthcare
- Financial data analysis
- Telecommunication Industry
- Biological Data Analysis
- Retail Industry
- Other Scientific Applications

Medical / Pharma

Computer Assisted Diagnosis
Characterization/prediction of patient's response to product correct dosages
Identification of successful medical therapies.
Study of relationships between dosage and potentially related adverse events
Insurance and Health Care
Discovery of medical procedures that are claimed together through claims analysis
Identification of customers, like who are potential buyers for new policies.
Detection of behavior patterns capable of identifying risky customers.
Detection of fraudulent behavior in polices and claiming activities.

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few common cases:

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classifying and clustering the customers for targeted marketing schemes.
- Detection of money laundering and other type of financial crimes.

Telecommunication Industry

Today the Telecommunication industry is one of the most emerging industry. Which provides services like fax, pager, cellular phone, Internet messenger, images, e mail, e chat, web data transmission etc.Due to the development of computer and communication technologies, the telecommunication industry is rapidly spreading out. This is the major reason behind data mining plays a very important role in understanding the business in proper way.

Telecommunication industry

In Telecommunication industry data mining helps identifying the telecommunication patterns, find fraudulent activities, make better use of resource, and improve quality of worshipping in various areas. Here is the list examples for data mining, which improve telecommunication services:

- Multidimensional Analysis of Telecommunication data.
- Pattern analysis of Fraudulent activity
- Unusual pattern Identification
- Multidimensional association and linear patterns analysis.
- Telecommunication services in mobile.
- Uses of visualization tools in telecommunication data analysis.

Biological Data Analysis

Now a day we used to see that, there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research. In Bioinformatics, Biological data mining is a very important part. Following the aspects of Data mining which contribute for biological data analysis in vast:

- Semantic integration of heterogeneous , distributed genetics that applies recombinant in DNA (genomic) and (proteomic)a study of proteins and their structure, functions databases.
- Alignment, indexing , similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns in genetic networks and protein pathways.
- Analysis of Association and path.
- Visualization tools in genetic data analysis.

Retail Industry

Data Mining has its great application in Retail Industry because it collects vast amount of data from sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of collected data will continue to expand rapidly because of increasing ease, availability and popularity of web.

The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improve the quality of customer service and good customer possession and satisfaction. Here are some examples of data mining in retail industry:

- Design and Construction of data warehouses based benefits on data mining.
- Multi scope analysis of sales, customers, products, time and region.
- Customer possession.
- Product recommendation and cross-referencing of items.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data groups, which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data being gathered in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics. Data mining in field of Scientific Applications:

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Domain and Visualization specific knowledge.

IV. SIX OF THE BEST OPEN SOURCE DATA MINING TOOLS:

Rapid Miner (formerly known as YALE)

Written in the Java Programming language, template-based frameworks are done by using rapid miner. Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools.

Rapid Miner provides functionality data preprocessing and visualization, predictive analytics and statistical modeling, in addition of data mining. WEKA and R scripts makes even more powerful learning schemes, models and algorithms.

Rapid Miner is distributed under the AGPL open source license and can be downloaded from the number one business analytics software **Source Forge**.

Orange

Python is getting in popularity because it's simple and easy to learn. When it comes to looking for a tool for your work and if you are a Python developer, look there is no further option than Orange, a Python-based, powerful and open source tool for both learner and experts in particular domain.

You will get addict to this tool's while visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with copy of features for data analytics.

NLTK

Comes to language processing tasks, nothing can beat NLTK. NLTK provides a pool of language processing tools. It includes data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks. All we need to do is install NLTK, pull a package for your favourite task and you are ready to go. Because it's written in Python, we can build applications on top of it, customizing it into small tasks.

WEKA

WEKA was developed for analyzing data from the agricultural domain. WEKA is an original non java version. With the Java-based version, the tool is used in many different applications. Which includes visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License, It is a big plus compared to Rapid Miner, because users can customize it according to their requirement.

WEKA supports several standard data mining tasks those are: data preprocessing, clustering, classification, regression, and feature selection.

R-Programming

Project R, a GNU project, is written in R programming. In no particular order It is written in C and FORTRAN.For statistical computing and graphics **R programming** acts as a free software programming language and software environment. The R language is widely used among data developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity and substantially in recent years. Besides data mining provides statistical and graphical techniques, including sequence and non sequence modeling, attic statistical tasks, time-series analysis, classification, clustering, and others.

V. CONCLUSION

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships which can identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. That is why I am choose data mining. With the help of these tools and techniques I will do my research in efficient way, and I am sure about my research will help the society in some way .Thanks to everyone who leads me for my project.

REFERENCES

[1] Alexander,D.(n.d)data mining.Retrieved from the universityof texasat Austin:college of liberal arts: http://www.laits.utexas.edu/~anorman/f

[2] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39, 1996.

[3] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[4] Elder, John F., IV and Daryl Pregibon, (1996), Advances in Knowledge Discovery & Data Mining, "A Statistical Perspective on KDD."

[5] Chen, M. et al, 1996. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 8.

[6] Han, J., & Kamber, M. (2006). *Data mining concepts and techniques.* Boston: Elsevier.

[7] Dunham, M.H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

BIOGRAPHY

Mr. V.S.Jagadheeswaren, working as a Assistant professor, Department of Information Technology, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India.

Miss.V.N.Saranya, Pursuing M.Phil Research Scholar, Department of Computer Science, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India.