ISSN: 2321-7782 (Online)

Volume 2, Issue 3, March 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Paper / Case Study

Available online at: www.ijarcsms.com

History and Current and Future trends of Data mining Techniques

Monika D. Khatri[1]
Dept. of Computer Engineering
Sipna College of Engg. and Technology
Amravati, Maharashtra – India

S. Dhande [2]
Assistant Professor
Dept. of Computer Engineering
Sipna College of Engg. and Technology
Amravati, Maharashtra – India

Abstract: ***We are surrounded with data. From human life birth people have been seeking patterns in data. We have grown accustomed gradually to the fact that there are tremendous volumes of data filing our computers, networks and lives. In all sectors such as government agencies, scientific institutions, and business have all dedicated enormous resources to collecting and storing data. From this large data only a small amount of these data will ever be used while others will be not useful for the task. There is need to understand these large data. The ability to extract useful knowledge hidden in these data is called as data mining and to act on that knowledge is becoming increasingly important in today's competitive world. This paper represents the review of various data mining techniques.***

I. INTRODUCTION

Data mining has become an established discipline within the scope of computer science. Data mining came into use late 80s within the research community [3]. By the early 1990s data mining was commonly recognized as a sub-process within a larger process called Knowledge Discovery in Databases or KDD. Other sub-processes that form part of the KDD process are data preparation and the analysis/visualization of results. The popularity of data mining increased significantly in the 1990s , notably with the establishment of a number of dedicated conferences such as the ACM SIGKDD annual conference in 1995,and the Euopean PKDD and the Pacific/Asia PAKDD conferences in 1997[3].The current trend in data mining is big data. The data minning is a capability of extracting data from a large set with respect to its volume, complexity. The popularity of data mining has continued to grow over the last decade with a particular current emphasis on mining non-standard data (i.e. non-tabular data).

II. KNOWLEDGE DISCOVERY PROCESS

The knowledge process is described as follows:

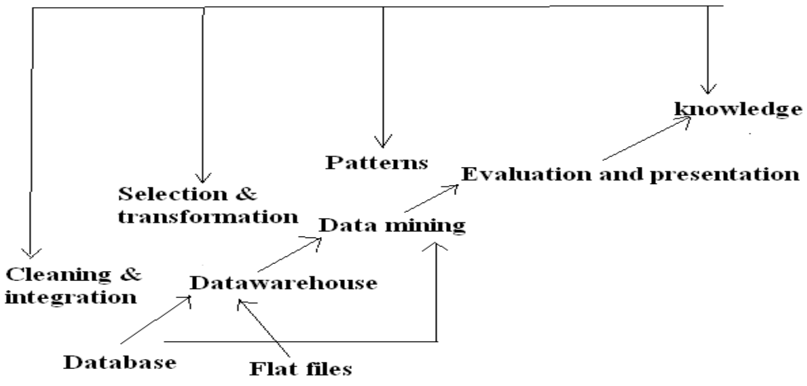

Fig.1 Data mining process.

The various processes are:

- Data cleaning: Remove noise that is unwanted data.
- Data Integration: Integration means combining multiple data sources.
- Data selection: Select related data to task from database.
- Data transformation: Convert the data into appropriate form that will be easy to mine.
- Data mining: a process such as association, regression, classification to extract data patterns.
- Pattern Evaluation: Evaluate the output of data mining process and identify the interesting measures.
- Knowledge Representation: Various techniques are used to present the mined data to the user [4].

III. DATA MINING TECHNIQUES

There are several data mining techniques that have been developed such as association, classification, clustering, prediction and sequential patterns, etc., are used for knowledge discovery from databases.

A. Association

Association is a technique which is used to find a pattern that is based on a relationship of a particular item on other items in the same operation [5]. The AIS that is Agrawal, Imielinski, Swami Algorithms was the first algorithm proposed for mining association rule [5].It focus on improving the quality of databases and decision support queries. In this algorithm only one item for association, for example we only generate rules like $a \wedge b \Rightarrow c$ but not those rules as $a \Rightarrow b \vee c$. Apriori is a great improvement in the history of association rule mining which was first proposed by Agrawal [6].The AIS requires many passes over the database which generate many candidate item sets and store counters of each candidate .Apriori-TID and Apriori-Hybrid [6][7] are modifications of the Apriori algorithm. Disadvantages of this algorithm are that:

- ✓ Complex process generation that uses space, time and complexity.
- ✓ Multiple scan of database

To break the two bottlenecks of Apriori series algorithms, association rule mining using tree structure have been designed such as FP-Tree in Han et al.2000 [5] and frequent pattern mining.

B. Classification

Classification is used to classify each item in a set of data item into one of predefined set of classes or groups. The desired classifiers can take many forms: decision trees, Support Vector Machines. The most influential decision tree generation algorithm with respect to data mining is Quinlan's C4.5 algorithm which came into notice 1993[8]. The most frequently referenced classification ARM algorithm is CBA algorithm [9]. other notable classification techniques include regression, for example the Cart algorithm [10],and Naïve Bayes[11].An artificial neural network is an interconnected group of artificial neurons that uses mathematical or computational model for information processing in Freeman et al,1991.In 1949,Donald Hebb pointed out the fact that neural pathways are strengthened each time they are used. In 1982 the concept was modified by John Hopfield. In 1986, with multilayered neural networks appeared. Neural networks are applied to data mining in 1997 in Craven and Sahvlik.

C. Clustering

Clustering is the process of organizing objects into groups whose members are similar by any means. A cluster is a collection of objects which have similarity between them and are dissimilar to the objects belonging to other clusters [12]. Clustering has always been used in statistics [13] and science [14].The introduction into pattern recognition [15] such as speech and character recognition. Machine learning clustering algorithms were applied to image segmentation and computer

[16].Statistical approaches to pattern recognition [17] and [18]. Clustering is also widely used for data compression in image processing, which is also known as vector quantization [19]. Clustering in data mining was brought to life by intense developments in information retrieval and text mining [20];[21];[22],spatial database applications [23],sequence and heterogenous data analysis[24],Web applications in [25];[26];[27],DNA analysis in computational biology[28],and many others.

D. *Prediction*

It is one of the data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables. In data mining, independent variables are attributes already known and response variables are what we want to predict [29]. The task of prediction is use to predict the future value from past data. In order to do this, one needs to build a predictive model for the data. The autoregressive models, example, can be used to predict a future value as a linear combination of earlier sample values, provided the time series is assumed to be stationary [30], [31] and Hastie et al 2001. Another popular work-around for non stationarity is to assume that the time series is piece-wise (or locally) stationary. The series is then broken down into smaller "frames" into which the stationarity condition can be assumed to hold and then separate models are learnt for each frame. For example, neural networks have been put to good use for nonlinear modelling of time series data in Sutton 1988,[32], [33], [34].The prediction problem for symbolic sequences has been addressed in AI research.

IV. Current and Future Trends

A. *Current trends[34]*

The following are the current trends of data mining:

1) *Fight against terrorism:*

Terrorism is a crime which is nowadays increased. The maior attack was 0-11 attack, almost all countries implemented new laws to fight against terrorism. Various programs were launched against the attack but they faced problems:

- Database of such programs contained text, video, audio, image, etc.
- The execution time increased as the size of data increased.

Example: 230 cameras were placed in xyz country to note the vehicles plate number. So, the estimation said that 40,000 vehicles pass camera every hour that means the camera should note 10 vehicles per second which resulted in heavy load on both hardware and software.

2) *Bio-informatics and cure of diseases:*

Health is the most important part of human life. If health is well than all is well. So, data mining can be used to cure diseases. The recent survey showed that large number of people is dying due to heart attacks, cancer, HIV, etc. The data mining techniques are used to predict the diseases from the past and current data of a patient.

3) *Web and semantic web:*

Nowadays internet has become the most important and necessary thing of human life. All the work is done through internet. But web contains lot of data i.e unstructured. Data mining can organize them called as semantic web. Social networking sites such as facebook are using FOAF technology for tagging. It is serving a lot in web.

4) *Business Trends:*

Today, business needs are larger but faster and accurate. Data mining technique such as prediction and classification are used in business. Example: Stock prediction is possible using data mining technique. Data mining vto improve the business as well as productivity of it.

B. *Future trends[35]*

The following are the future trends:

1) *Distributed/collective data mining:*

Today data mining works on database or dataware house in which data is located in one place. Future scope is that data mining will be used to mine the data that are located in different places which is known as distributed data mining (DDM).Therefore, the goal is to mine distributed data which is located in heterogeneous sites. The data is first analysised at local level then the local data from all the sites are combined to form global level data. The problem is that the data in different sites may have different characteristics and they may become ambiguous.

2) *Multimedia mining:*

Multimedia includes text, animation, audio, video and images. Multimedia data is different from simple data. To held multimedia data we can create a data cube which can be used to convert the multimedia data into a form that can be used to apply data mining techniques but considering the characteristics like shape, color, dimensions, etc. Another data is audio data which can also be mined. The logic is to use audio signals to indicate patterns of data.

3) *Spatail and geographic data mining:*

Spatial and geographic data contains data such as natural resources, orbiting satellities and spacecraft which transmit images of earth. Mostly such kind of data is in images. S partial data are the data which are topological information and distance information. The challenges in building spatial datawarehouse are the collection of data from heterogeneous sources.The spatial data cubes is created which is part of spatial datawarehouse.

4) *Phenomenal data mining:*

It focuses on the relationship between phenomena and data. Ex: The receipt of purchase from supermarket of a customer can reveal many things about the customers such as age, purchasing habits, etc. These could be achieved either by implementing a program or put in a database which examines data for phenomena. The major challenge of building such warehouse is a coding part of common sense which is found to be quite difficult.

V. Conclusion

In this paper we have briefly reviewed various data mining techniques history and future trends of data mining. This will help researchers to know about the inventions and future of data mining.

References

1. Data mining-know it all by Soumen Chakrabarti ,Earl Cox Eibe Frank, Ralf Hartmut Güting, Jaiwei Han Xia Jiang, Micheline Kamber, Sam S. Lightstone Thomas P. Nadeau, Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten.
2. Introduction to Data mining Techniques by Dr.Rajni Jain.
3. The Knowledge Engineering Review,vol.00.0,1-24.2004,Cambridge University Press DOI:10.1017/S00000000000000,Data Mining: Past, Present and Future by FRANS COENEN Department of Computer Science, The University of Liverpool, Liverpool, L693BX, UK
4. Data mining: concepts and techniques second edition,Jiawei Hn,University of lions at Urbana Champaign,Micheline Kamber.
5. Association Rule Mining:A Survey by Qiankun Zhao Nanyang Technological University, Singapore and Sourav S. Bhowmick Nanyang Technological University, Singapore.
6. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207-216.
7. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 487-499.
8. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.Yan, X. and Han, J. (2002). gSpan: Graph-Based Substructure Pattern Mining. Proc. IEEE International Conference on Data Mining (ICDM '02), IEEE, pp721-724

9. Liu, B., Hsu, W. and Ma, Y. M. (1998). Integrating classi_cation and association rule mining. Proc KDD-98, ACM press, pp.80-86.
10. Breiman, L., Friedman, Y., Olshen, R. and Stone, C. (1984). Classi_cation and Regression Trees. Wadsworth, Belmont, CA, 1984
11. Hand, D.J. and Yu, K. (2001). Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, pp385- 398.
12. Survey of Clustering Data Mining Techniques by Pavel Berkhin,Accrue Software,Inc.
13. ARABIE, P. and HUBERT, L.J. 1996. An overview of combinatorial data analysis, in: Arabie, P., Hubert, L.J., and Soete, G.D. (Eds.) Clustering and Classification, 5-63
14. MASSART, D. and KAUFMAN, L. 1983. The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis. John Wiley & Sons, New York, NY.
15. DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.
16. JAIN, A.K. and FLYNN, P.J. 1966. Image segmentation using clustering. In Advances in Image Understanding: A Festschrift for Azriel Rosenfeld, IEEE Press, 65-83.
17. DEMPSTER, A., LAIRD, N., and RUBIN, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1, 1-38
18. FUKUNAGA, K. 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, CA
19. GERSHO, A. and GRAY, R. M. 1992. Vector Quantization and Signal Compression. Communications and Information Theory. Kluwer Academic Publishers, Norwell, MA.
20. CUTTING, D., KARGER, D., PEDERSEN, J., and TUKEY, J. 1992. Scatter/gather: a cluster based approach to browsing large document collection. In Proceedings of the 15th ACM SIGIR Conference, 318-329, Copenhagen, Denmark
21. STEINBACH, M., KARYPIS, G., and KUMAR, V. 2000. A comparison of document clustering techniques. 6th ACM SIGKDD, World Text Mining Conference, Boston, MA.
22. DHILLON, I. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In Proceedings of the 7th ACM SIGKDD, 269-274, San Francisco, CA.
23. XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J. 1998. A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th ICDE, 324-331, Orlando, FL.
24. CADEZ, I., SMYTH, P., and MANNILA, H. 2001. Probabilistic modeling of transactional data with applications to profiling, Visualization, and Prediction, In Proceedings of the 7th ACM SIGKDD, 37-46, San Francisco, CA.
25. COOLEY, R., MOBASHER, B., and SRIVASTAVA, J. 1999. Data preparation for mining world wide web browsing. Journal of Knowledge Information Systems, 1, 1, 5-32.
26. HEER, J. and CHI, E. 2001. Identification of Web user traffic composition using multimodal clustering and information scent. 1st SIAM ICDM, Workshop on Web Mining.
27. FOSS, A., WANG, W., and ZAANE, O. 2001. A non-parametric approach to Web log analysis. 1st SIAM ICDM, Workshop on Web Mining, 41-50, Chicago, IL.
28. BEN-DOR, A. and YAKHINI, Z. 1999. Clustering gene expression patterns. In Proceedings of the 3rd Annual nternational Conference on Computational Molecular Biology (RECOMB 99), 11-14, Lyon, France.
29. www.wikipedia.com,prediction.
30. Box G E P, Jenkins GM, Reinsel G C 1994 Time series analysis: Forecasting and control (Singapore:Pearson Education Inc.)
31. Chatfield C 1996 The analysis of time series 5th edn (New York, NY: Chapman and Hall)
32. Haykin S 1992 Neural networks: A comprehensive foundation (New York: Macmillan)
33. Koskela T, Lehtokangas M, Saarinen J, Kaski K 1996 Time series prediction with multilayer perceptron, FIR and Elman neural networks. In Proc. World Congress on Neural Networks, pp 491–496
34. Rise of data mining:Current and future application areas by Dharminder Kumar 1Professor and Dean, Faculty of Engineering & Technology, Guru Jambheshwar University of Science & Technology, Hisar Deepak Bhardwaj Research Scholar, Department of Computer Science & Engineering Guru Jambheshwar University of Science & Technology, Hisar in IJCSI,issues vol.8,issue 5,no.1,September 2011,ISSN(online):1694-0814.
35. Data mining:future trends and applications by Anman Naidu Paidi,Asst professor of CSE centurion University,Odisha,india,International journal of modern engineering research(IJMER),vol 2,issue 6,nov-dec 2012.pp-4657-4663.ISN:2249-6645

AUTHOR(S) PROFILE

Monika Khatri received B.Tech. degree in Information Technology in 2013 from Government College of Engineering Amravati(An Autonomous Institute of Government of Maharashtra) and now pursuing M.E. from Sipna College of Engineering and Technology, Amravati (under Sant Gadge Baba Amravati University).