International Journal of Science and Advanced Innovative Research
ISSN: 2536-7315 (Print) 2536-7323 (Online)
Volume 5, Number 2, June 2020
http://www.casirmediapublishing.com

DATA MINING: THEORY, CONCEPT AND TECHNIQUES

Mustapha Ibrahim[1] and David O. Tayo[2]
[1&2] Department of Computer Science
Federal Polytechnic, Damaturu. Yobe State
Email: mustee2004@yahoo.com or babakhalil2012@gmail.com

ABSTRACT
Organizations recently developed transaction processing technology that requires data captures in large amount and match the speed of processing of the data into information which can be utilized in making decision. Data mining, the extraction of hidden predictive information from large databases, is a newly powerful technology with great potential that help organisations to project on the vital information in their data warehouses. Machine learning is used for statistical and visualization techniques to discover and present knowledge in a form which is easily intelligible to humans. Data mining tools are used to predict future trends and behaviours, allow businesses to make proactive, knowledge-driven decisions. Most organisations received and store large volume of data about their businesses and most of these data are not used to analyse useful information form it due to inability to derive a viable information form it. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining tools support organisation to ascertain valuable information from the data set. To deliver a large volume of data in term of speed and accurate the use of data mining tools becomes paramount for effective and efficient useful of information of future prediction. Financial organisations in these days make use of computer as a tool for adequate and effective storage of information or data.
Key words: Data Mining, Data warehousing, machine learning and Tools and Techniques

INTRODUCTION
As the world turned to global village in terms of communication and storing data, data mining becomes a vital tool for storing data or information securely and efficiently for feature retrieval. The use of data within organisation has been rapidly increase in terms of capacity and process in daily basis, in other to safe guide the data for effectively and efficiently, the concept of data mining tool and techniques has to be involved so as to allow a user's to extract the information. Reorane and Kulkarni (2011) are of the opinion that, Data mining becomes essential tool for daily activities within an organisation that can be used to store and analyse data of such organization based on user communication and financial transaction. Considering the fact that large amount of data are been collected, process and transmit for the purpose of achieving goal, there is equally a need to transmit the data in a secure and efficient manner so as to get the use out of it, using computerized system or information technology infrastructure. Huang, Liu and Chang (2012) agreed that, automated system can be used to enable organisation to receive and stored very large volume of information, which decision can be made from it. The report examine the concept of Data mining, Data mining tools and techniques, Challenges and Application of Data mining, Limitation of Data mining and draw a conclusion.

Concept of Data Mining

In the modern era, data mining become an indispensable tool for financial organization to transformed data into meaningful information for better decision making to achieve a viable advantage over competitors. Data mining is the process of investing and examining of huge amounts of data with a specific end goal to find significant patterns and tenets that can enhance business choice, for making decision in an organization (Wang, Lin and Hou, 2015). Raorane and Kulkarni (2011) and Lone and Khan (2014) are of the view that, Data mining is a methodology that is used to collect and analyse data form different viewpoint, in other to get meaningful information out of it, which increase the productivity of an organizational goal and objectives. Furthermore, in a relational database which is very large in size, data mining can be used to gather information using different patterns among the number of fields within the relational data bases. Lone and Khan (2014) added that, Data mining can be described as the innovation which joins the measurable techniques (statistical techniques) and scientific comparisons (Mathematics equations) that are utilized as a part of an endeavour to distinguish the significant relationships between variables in the Historical data, to gauge or perform investigation on the data; or focus any huge relationship inside the data been collected. With advent of Data mining principals, strategic level with in an organization find it simple to decide on issues that are affecting the organisation goals and also addresses problems relating to their financial status (Kaur and Aggarwal, 2010 in Mohammed et al, 2013). Naidu and Tiwari (2014) and Agrawal (2013) view data mining as a subfield of computer science which process, transmit and discover patterns in a large volume of data sets. Furthermore, they added that data set may intersect or integrate with other branches of computer science such as machine learning, artificial intelligent, data base and statistical for data analysis. Agrawal (2013) added some of the subfield in computing to can be integrated with data mining, such as pattern recognition, neural networks, data visualization, information retrieval, image and signal processing for information processing and retrieval. It is an effective new innovation to help organizations concentrate on the most critical data in their data warehousing. Naidu and Tiwari (2014) and Mohammed et al (2013) agreed that, the main goal of the data mining is to derive a useful information from a data set and change it into a reasonable pattern for future decision making in an organization.

Data warehouse, which includes data cleaning and data integration can be seen as a vital pre-processing venture for data mining. In any case, a data warehouse is not a prerequisite for data mining. Building a vast data warehouse that combines data from different sources, solve the problem and issues arise in data integrity, and loads the data into a database. Data mining uses the data warehouse as the wellspring of Information for Knowledge data discovery (KDD) frameworks through an amalgam of artificial intelligent and statistical related procedures to discover affiliations, groupings, orders, groups and future forecasts (Agrawal, 2013).

International Journal of Science and Advanced Innovative Research
ISSN: 2536-7315 (Print) 2536-7323 (Online)
Volume 5, Number 1, March, 2020
http://www.casirmediapublishing.com

Data Mining Tools and Techniques

Data mining tools forecast future challenges and performances, and allowed organisation to make a predictive decision based on foreseeing event. For a financial organisation, to discover a previous unknown statistical patterns can offer valuable a well-planned solution for proper function of their organizational environment. On the other hand, Data-mining techniques are basically divided into two aspect: predictive method and descriptive method (Naidu, Tiwari, 2014) and (Agrawal, 2013). Agrawal (2013) further explain predictive method as, a model that uses statistical method (Regression, time series analysis) to predict data using unknown parameters; whereas the descriptive model identified the relationship that exist within it and explore the difference and similarity such as cluster, Association Rule and Sequence discovery.

To deliver a large volume of data in term of speed and accurate the use of data mining tools becomes paramount for effective and efficient useful of information of future prediction. Financial organisation currently make use of computer as a tool for adequate and effective storage of information or data. Mohammed et al (2013) and Naidu and Tiwari (2014) opined that most organisations received and store large volume of data about their businesses and most of these data are not used to analysed useful information form it due to inability to derive a viable information form it. Data mining tools support organisation to ascertain valuable information from the data set. According to Ramamohan et al (2012) classified data mining tools into three (3) parts, each of which has his own pros and cons depending on the context that the organisation want to make use of it. There are as follows: Traditional Data Mining, Text-Mining and Dash boards mining tools.

Traditional Data Mining Tools: Traditional Data mining Program help organizations make use of data trends and patterns by utilizing various complex Algorithms and strategies. Some of these devices are introduced on the desktop to monitor the information and highlight patterns and others stored data which are not previously in the database. This tools is also available in both Windows and UNIX kernels or operating system, even though some are been design for a particular or specific operating system only. Furthermore, while some may focus on one database type, most will have the capacity to handle any information utilizing online expository transforming or a comparable innovation (Ramamohan et al, 2012) and (Naidu and Tiwari, 2014).

Dash Board: Dash board mining tool is a piece of software install in computer to observe the follow of information in the repository of the Database, for any slide change for the data or information in the dataset it usually reflect the changes in the dash board. The updated information are normally in form of chart or in tabular form, which draw the attention of organisation to see how well their performing. The information can also be used to predict future performance using the historical data. The dashboard system, becomes easy to use and also present information to the

strategic level to make decisions for the future performance of the company (Ramamohan et al, 2012) and (Naidu and Tiwari, 2014).

Text Mining Tool: Text-mining tool is a program that has the capacity to change data into different form or convert different kinds of text. It can mine a word text (Microsoft Word, Acrobat PDF documents) to simple text files. Information Content can be scan and converted into any kind of document format that is compatible with the database tool, with these user has a convenient and efficient way to gain access to a dataset without need to install different another application or open with another application. Getting these inputted information can provide organizations with an adequate information that can be extracted to discover new challenges, ideas and approaches (Ramamohan et al, 2012) and (Naidu and Tiwari, 2014).
In contrast Chen,Sakaguchi and Frolick (2000) are of the viewed that, in other to deliver an information in an organisation within a short possible period of time and of large volume, the need of data mining tools becomes necessary for proper solution to the problem. Furthermore, they classified Data mining tools into four (4) categories: Data, Hardware, Software and Network.

Data: Data in Data warehouses has four basic factors that enhance the productivity of data mining so as to eliminate the problems of computational intensive: these include data Integration which structure data for mining to become easy; Detailed summarised data which prevent for repeating a work and discovery of trends and patterns of data; Historical data which also organisation to focus on unforeseen event in their organisational cycle and lastly, the meta data which help in proving the actual data or a guide that decision could be drawn from by the end users in data mining (Chen, Sakaguchi and Frolick, 2000).

Hardware: Chosen Hardware become a paramount, since the greatest problem of data mining techniques become rigorous in term of computation. Database management system becomes powerful tools that can be used to mine data, most large Application of data mining require complex hardware components (Chen, Sakaguchi and Frolick, 2000).

Software: For an organisation to achieve its objective for proper decision making efficiently and effectively, data mining software need to become versatile so as the manager can view the information content differently. The versatility of software in data mining, is that the application should be able to display the content of the information in visualisation such as in graphs and tables (Chen, Sakaguchi and Frolick, 2000).

Network: Network becomes essential due to the increase in the development and implementation of client/server data mining in the recent era. Moreover, the use of network could be better which has the capacity to tackle net traffic load and enhance the organisational response rate (Chen,

International Journal of Science and Advanced Innovative Research
ISSN: 2536-7315 (Print) 2536-7323 (Online)
Volume 5, Number 1, March, 2020
http://www.casirmediapublishing.com

Sakaguchi and Frolick, 2000). Liao, Chu and Hsiao (2012) identifies various techniques that are available in data mining, these techniques defer for one type to another and are used to solve different real life application; such as: statistics, neural networks, decision trees, genetic algorithms, and visualization techniques. Below are some of main categories of data mining techniques which can further divide into specific application.

Decision Tree Analysis: Decision tree is one of the commonly used Data mining techniques that is used to show the flow of information in a hierarchical structure (Liao, chu and Hsiao, 2012) and (Naidu and Tiwan, 2014). Wang, Lin and Hou (2015) opined that, the advantage of decision tree over the others is that, it uses conditional statement (IF-THEN) rule to show the flow of outer information.Weiping and Yuming (2013) are of the viewed that, the flow of information in decision tree are sequential which can be easily understand.

Neural Network: Is an Artificial Intelligent system, which is been divide into two (2) parts: the Connector and the Neuron (Wang, Lin and Hou, 2015). The Neuron process the information required and the connector are communicating channels that allow the flow of information between neurons (Wang, Lin and Hou, 2015) and (Liao, chu and Hsiao, 2012).

Logistic Regression: Is a Statistical classification that uses probability theory to divide a set of data into different classes. The logistic regression model measure the relationship that exist between dependent and independent variables using values predicted in the dependent variable (Wang, Lin and Hou, 2015). Naiwa and Tawari (2014) viewed the overall model as Regression model, which is used to analyse data by using statistical Hypothesis to agree or disagree on setting Issues or result.

Cluster Analysis: the cluster analysis is a statistical method that groups a task into different segment or unit, so that each object in a unit are not similar to other object in different unit (Wang, Lin and Hou, 2015) and (Naiwa and Tawari, 2014). However, Agrawal (2013) added that the cluster analysis becomes the most expenses, since group can be used as pre-processing approach for each classification and selection of a subset of attribute.

Data Mining Issues and Challenges

As data mining enterprises continue to change, several reason are considered when chosen and implementing a technique. The basic challenging issue in data mining is chosen an appropriate data mining technique in an appropriate problem context. Agrawal (2013) figure out some issue that are consider selecting and implementing a data mining technique so as to avoid an oversight. However, the challenging issues are: Data Quality, Interoperability, Mission creep and privacy.

Data Quality: Data quality is one of the greatest issues or challenges of data mining. For data to be accepted it has to be accurate and complete. Data quality can likewise be influenced by the structure and consistency of the Data being examined or analysed (Agrawal, 2013).

Interoperability: Refers to the capacity of a computer system to work with other systems that are having the same process or standard (Agrawal, 2013).

Mission Creep: Mission creep is one of the prominent risks factor of data mining and symbolises how control over user's information can be insubstantial proposition. Furthermore, Mission Creep allow data to be used for another purpose rather than what was originally meant for (Agrawal, 2013).

Privacy: Has an extra information offering and data mining activities have been declared, expanded consideration has concentrated on the ramifications for protection and privacy. Security and privacy focus both on real tasks proposed, and also worries about the potential for data mining applications to be extended outside their purposes which are meant for (Chen, Sakaguchi and Frolick ,2000) and (Agrawal, 2013).

Application of Data Mining

Data mining is applicable in also every aspect of Human Endeavour, which has the ability to store, process and retrieved information or data as the need arise. Different literature have discuss numerous application of data mining. Below are some of the basic application of data mining:

Retail: Retails uses data mining concept to predict customers' buying behaviour and make reference to it in future (Wang, Lin and Hau, 2012) and (Chen, Sakaguchi and Frolick, 2000). Some basic Retail applications of data mining in term of performance include: performing basket analysis (affinity analysis), database marketing, sales forecasting, Merchandise planning and allocation (Rygielski, Wang and Yen, 2002).

Telecommunication: Telecommunication industries around the globe face raising in the state of rivalry which is compelling them to forcefully advertise uncommon evaluating projects went for holding existing clients and attract new ones (Rygielski, Wang and Yen, 2002).

Banking/Finance: Financial organisations utilises data mining for various application such as card marketing, card holder pricing and profitability, fraud detection, predictive life-cycle management (Agrawal, 2013) and (Rygielski, Wang and Yen, 2002). Data mining task is used in constructing credit scoring models from a credit database (Huang, chen and Wang, 2007). Chen, Sakaguchi and Frolick (2000) pointed that one of the successful bank that uses data mining in credit card industry is American Express and Citibank respectively.

International Journal of Science and Advanced Innovative Research
ISSN: 2536-7315 (Print) 2536-7323 (Online)
Volume 5, Number 1, March, 2020
http://www.casirmediapublishing.com

Manufacturing: Manufacturers uses data mining to predict futures trends and also meet with what customer requirement in other to enhance the profitability of their organisation. It empowers producers to foresee number of clients who will submit warranty claims which will be calculated (Ryzielski, Wang and Yen, 2002).

Airlines: Airline industry uses Data mining to enables understand and make decision of what their customers' need and increase their route services, since competition is imperative in the air industry (Chen, Sakaguchi and Frolick, 2000).

Medicine: In medicine Data mining plays a vital role of analysing and interpretation of patient's record as well as used to describe patient behaviour toward a particular disease, it also measure how often surgery are undertaking (Dudley, 2009).

The Trends of Data Mining Development in Current and Future Research Domain

Data mining is another sort of intelligent information processing techniques. With the fast improvement of information technology, the application in the field of data mining will increase and extend interminably, particularly in the military, security, business intelligence applications (Weiping and Yuming, 2013). Furthermore, Data mining is specifically confronting huge databases, so the algorithm of Data mining has to be efficient and effective for it to achieve its purpose. The key trend of data mining in current and future research domain is to analysed complex data, as well as finding a suitable technique that will handle the problem context. Khatri and Dhande (2014) opined that, that Big data is the current trend of data mining, which is capable of extracting data from a dataset in respective of it complexity and volume.

The following are the current and future trends in data mining respectively:

Current Trends

Fighting against Terrorism: As the whole world is facing with problems of terrorism, which increases in daily bases; several laws where been enacted to fight and control terrorist attack. Various software's were install to monitor terrorism attack but failed due to problems of mixture of data contain such as text, video, audio and image. This is as a result of increase of execution time and as well as the size of the data (Khatri and Dhande, 2014).

Bioinformatics and Cure of Diseases: Khatri and Dhande (2014) and Goele and Chanana (2012) agreed that, data mining can be used to cure a disease based on prediction of the current and past historical event.

Web and Semantic Web: The use of web has become part human being day to day active. Because most or almost all are work are done through the use of internet, but yet the data it contained is

unstructured or having different syntax (Goele and Chanana, 2012). Data Mining can be structured be using a semantic web (Khatri and Dhande, 2014).

Business Trends: As organization need large repository to stored and retrieved data fast and accurate, the prediction and classification techniques are used to achieve the productivity and as well improve the organizational goals (Khatri and Dhande, 2014) and (Goele and Chanana, 2012).

Distributed/ Collective Data Mining: The concept of distributed data mining is to mine a data that are of different location to achieve purpose. The challenge is that, the data may be ambiguous since are coming from different location (Khatri and Dhande, 2014).

Multimedia data mining: In multimedia data a cube is created, which are used to convert multimedia data into a separate form so that mining technique can be applied, taking into consideration the shape, color and dimensions cube (Khatri and Dhande, 2014).

Spatial and geographic data mining: In a spatial and geographic data mining it hold images in form of data such as Natural resource, orbit satellite and spacecraft which show images of earth based on latitude and longitude (Khatri and Dhande, 2014).

Phenomenal data mining: Khatri and Dhande (2014) viewed it as a relationship that exists between data and phenomena which can be examined using data mining technique. Furthermore, challenging facing future trend is the coding aspect with remain so difficult.

Limitation of Data Mining

Data mining requires talented specialized in data mining who can structure and examine the analysis of data and translate the result that is made. Therefore, the limitation of Data mining are essential or personal related, as opposed to technology related. Data mining can help uncover patterns or trends and connections, it doesn't tell the client the worth or importance of these trends. Another limit factor of Data mining is that while it can distinguish associations in the middle of practices and/or variables, it doesn't essentially recognize a causal relationship (Agrawal, 2013).

CONCLUSION

The significant increase in the use of data mining for proper and better decision making with and outside organization, the Knowledge discovery becomes a key answer to decision making. Data mining includes the utilization of data Analysis tools to discover unknown issue for previous data, trends which are valid are compare with the relationships in a large repository system, Data mining is getting to be progressively normal in both the governmental and private sectors. Data mining is applicable in almost every firm that uses various data types, for efficient and effective research and development which seems to be a great challenge. Data mining tools and techniques help

in finding significant and important data from a very large data set. For successful utilization of mine technology, strategic level of management most understand the concept of data mining tools and techniques, so as to make choice of the best applicable tool and techniques that will suit their application. Organizations today are under tremendous pressure to strive in an environment within a short period and met with deadlines and as well minimizes profits. Business processes that require data to be extracted and manipulated prior to use will no longer be acceptable. Instead, enterprises need rapid decision support based on the analysis and forecasting of predictive behaviour. Data- warehousing and data-mining techniques provide this capability

REFERENCE

Agrawal, D. (2013) A Comprehensive Study of Data Mining and Application. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* [online], 2pp. 249-252

Chen, L., Sakaguchi, T. and Frolick, M. (2000) Data Mining Methods, Applications, and Tools. *Information Systems Management* [online], 17(1), pp. 1-6. Available at:<http://wlv.summon.serialssolutions.com>.

Dudley, C. (2009) *Database technologies*. Lecture 6: Data warehousing [online]. Available at: < http://wolf.wlv.ac.uk>.

Huang, T.C., Liu, C. and Chang, D. (2012) An empirical investigation of factors influencing the adoption of data mining tools. *International Journal of Information Management* [online], 32(3), pp. 257-270

Huang, T.C., Wu, I. and Chou, C. (2013) Investigating use continuance of data mining tools. *International Journal of Information Management* [online], 33(5), pp. 791-801

Khan, D.M., Mohamudally, N. and Babajee, D.K.R. (2013) A Unified Theoretical Framework for Data Mining. *Procedia Computer Science* [online], 17(0), pp. 104-113

Liao, S., Chu, P. and Hsiao, P. (2012) Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications* [online], 39(12), pp. 11303-11311

Lone, T.A. and Khan, R.A. (2014) Data Mining: Competitive Tool to Digital Library. *DESIDOC Journal of Library & Information Technology* [online], 34(5).

Mohammed, B., Mouhoub, M., Alanazi, E. and Sadaoui, S. (2013) Data Mining Techniques and Preference Learning in Recommender Systems. *Computer and Information Science* [online], 6(4), pp. 88

Naidu, H. and Tiwari, A. (2014) Data Mining and Data Warehousing. *International Journal of Engineering Sciences & Research Technology* [online], 3pp. 109-111

Ramamohan, Y., Vasantharao, K., Chakravarti, C.K. and A.S.K.Ratnam (2012) A Study of Data Mining Tools in Knowledge Discovery Process. *International Journal of Soft Computing & Engineering* [online], 2pp. 191-194

Raorane, A. and R.V.Kulkarni (2011) DATA MINING TECHNIQUES: A SOURCE FOR CONSUMER BEHAVIOR ANALYSIS. *International Journal of Database Management Systems* [online], 3pp. 45-56

Ryzielski, C., Wang, J. and Yen, D. C. (2002) Data Mining Techniques for Customer Relationship Management. *Technology in society* [online]. **24**(4), pp. 483-502.

Wang, J., Lin, Y. and Hou, S. (2015) A data mining approach for training evaluation in simulation-based training. *Computers & Industrial Engineering* [online], **80**(0), pp. 171-180