INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A LITERATURE REVIEW ON DATAMINING

Sankar K V[1], Dr.S.Uma[2], Subin P S[3], Ambat Vipin[4]

[1]PG scholar department of computer science and engineering HIT Coimbatore'
[2]Head of the department PG department of computer science and engineering HIT Coimbatore'
[3] PG scholar department of computer science and engineering HIT Coimbatore'
[4] PG scholar department of computer science and engineering HIT Coimbatore

PG DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
HINDUSTHAN INSTITUTE OF TECHNOLOGY
COIMBATORE, TAMILNADU, INDIA

Abstract-Data mining is one of the powerful new technologies that has emerged and It facilitates the users both individuals and organizations to dig and find data from a collection or a large cluster of data. Data mining is used to find patterns from a large group of data. For example an organization can find out customer behaviour from the data collected. The far reach of data mining extends to Artificial intelligence, business, education, scientific fields, machine learning etc. In this literature review a discussion of the basic concepts, applications, and the challenges in data mining is done.

Key Words*:* Data mining, Artificial intelligence, Information.

I. INTRODUCTION

Data mining [1] is the method of finding patterns, trends and relations by moving through a large amount of data stored in data repositories. This is done using techniques like statistics and mathematics. In data mining, data is analysed through different views in order to find the patterns that satisfy our needs. In short, data mining is also an analytical technique. Data is stored in various forms like images, text, sound, videos etc. Using data mining we can categorize data, relate similar data, and find the data occurrence patterns.

Another way of mentioning data mining is Knowledge Discovery in Databases also known as KDD [2]. It also involves extracting information and patterns from a large collection of data in a database. The main functionalities of data mining is to apply various methods to identify bits of information and to use this information for various decision making applications. For the last few years data mining and its applications in finding patterns and decision making, and has become an important breakthrough in the world of technology and it is now becoming one of the most sought out methods of data analysis.

Data mining is being applied in the fields of medicine, education, machine learning [3], pattern recognition, artificial intelligence etc.

II. DATA MINING

According to the Gartner Group data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques[4]
Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [5]. Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases [6].The data that is used to extract patterns in data mining is called as training data. Thus the training data is the data used to train a system to make decisions. Data mining has two ways in dealing with data namely Classification [10] and Prediction [10]

- ***Classification***: Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data.
- ***Prediction***: Predicts unknown or missing values

Data mining process can be considered as knowledge transformation from a fine grained space to a coarse grained space[11].
A new techinique called Ontology of core data mining entities exist by the name of OntoDMCore[12]. It defines the most essential data mining entities in a 3 layered ontological structure comprising of specification, implementation, and an application layer. It provides a representation al framework for the mining of structured data.
There are 2 methods of data mining for predicitng soil electeical resistivity namely Support vector machine(SVM)[22] and Least Squares Support Vector Machine(LSSVM)[22]

III. DATA MINING : A BRIEF HISTORY

Data mining was a term first introduced in 1990's and the technology roots back to a family of technologies namely machine learning, classical statistics and artificial intelligence.

- ***Machine learning*** is a combination of statistics and artificial intelligence (AI). It is considered that AI evolved to become machine learning because AI heuristics and statistical analysis are blended in machine learning. Machine learning is a technology that allows computers to understand and study the data they work on and make decisions based on what they have learned from those data. This is done by using statistics and adding AI heuristics to it.
- ***Statistics*** forms the base for most technologies that data mining is based upon. The examples include standard variance, regression analysis, cluster analysis, discriminate analysis, confidence intervals, standard deviation and distribution. All these help to analyze data patterns and their relations.
- ***Artificial intelligence*** uses heuristics to simulate the system, a behavior similar to humans which is the ability to think and make decisions on its own. This ability largely benefits data mining as it requires machines it make decisions without prompting.
  Data mining therefore is a combination of old and modern day statistics, Artificial intelligence and machine learning. These methods are combined to find patterns both hidden and within a large collection of data.

IV. USES OF DATA MINING

The primary usage of data mining is information extraction from a large number of data. Other uses include

- To obtain maximum information from a large amount of data that in normal provides only scanty information
- To interpret extracted data
- To find hidden patterns
- To use patterns for decision making

  Image mining [16] is a spin off form of data mining in which images are extracted from a large database of images to find specific patterns: Image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the image database. Image mining is more than just an extension of data mining to image domain. It is an Interdisciplinary endeavor that draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence.

V. FUNCTIONS OF DATA MINING

The main functions of data mining are given below[8]

- ***Classification***: Classification is finding models that analyze and classify a data item into several predefined classes.
- ***Sequencing***: Sequencing is similar to the association rule. The relationship exists over a period of time such as repeat visit to supermarket.
- ***Regression***: Regression is mapping a data item to a real-valued prediction variable.
- ***Clustering***: Clustering is identifying a finite set of categories or clusters to describe the data.
- ***Dependency Modeling***: Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables.
- ***Deviation Detection***: Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data.
- ***Summarization***: Summarization is finding a compact description for a subset of data.
- ***Data Cleaning***[20] removes noise from data,
- ***Data integration***[20] combines various data source
- ***Data Selection*** [20] transformation transforms data into the storm appropriate for mining.

VI. APPLICATIONS OF DATAMINING

Data mining is used in all fields of science and engineering. Some of the applications of data mining are listed below.

Medical/Pharmacy:

- Computer assisted diagnosis,
- Response to drug dosage,
- Successful prescription by study of patterns.

Insurance and health care:

- Finding medical procedure for claims through claims analysis
- Identifying potential buyers of drugs
- Identifying risky customers through behavior pattern analysis.

Banking:

- Fraud detection
- Risk management
- Hidden correlation discovery between financial indicators

The applications of data mining cited in some research papers are given below

- It is used in the field of forecasting [13] i.e. by bending the series data business gain can be achieved if data is converted in to information and then in to knowledge. This provides an ample opportunity to leverage numerous sources of time series data. Insurance frauds are also detected using data mining [19].

- Data mining is also used for Botnet detection [23]. A bot is autonomous software capable of performing its own functions, botnets are used in cybercrime as they are very powerful and can do denial of service, phishing, spamming and eavesdropping. Data mining and pattern detection of network traffic helps to prevent botnets from breaching security.

- Another field were data mining is used is in banks and financial institutions for credit scoring [14].

- Data mining is used in electronic commerce. Electronic commerce involves the use of information and communication technologies through internet platform. Data mining and web data mining techniques are used in the electronic commerce to understand customer patterns. [15]

- Medical data mining [17] is another important field where the large amount of data that is generated in various medical and health centers are mined to make the diagnosis, prognosis, and treatment schedules easier and quicker.

- It is used to extract application oriented models and patterns from a continuous, rapid flow of data streams in sensor networks [18].

- Pattern mining is used for activity recognition by searching for patterns of time segments in which same activity is performed or occurred [21].

- Data mining is also applied in social media to extract actionable patterns that can be beneficial for business, users, and consumers [24].

- Data mining is used in psychology [25] for pre diagnosis i.e. based on the data obtained from the diagnosis and treatment of various patients a pattern can be found for each mental disorder that is common to patients suffering from that particular disorder. Based on this pattern we can predict if a person is susceptible to an issue of the same type.

VII. DATA MINING CHALLENGES

Like all other technologies data mining also has its share of issues. The following are the issues in data mining.

- ***Security and social issues***

  It is well known that data mining mainly focuses on a large amount of data to find out the relevant data required. So security [8] is an issue in data mining. The profiling, pattern making and matching of various people and other sources will lead to a breach in the privacy of data on people, even it is susceptible to illegal accesses. Also creation of training data based on a user's behavior may cause breach of privacy for that individual and it can be harmful if those details are linked to other seriously confident data. Data mining poses a threat to usage of private and confidential information without control.

- ***User interface issues***

  Data that is obtained through data mining need to be presented in such a manner that the users must be able to analyze and understand the extracted data. The ability to see data clearly is entirely dependent on the tool used to present the data. A good visualization allows proper and accurate interpretation of data.

  However even though there are good visualization techniques available further research is necessary for avoiding problems like screen real estate and thus allow the user to focus on the data and refine the mining tasks and also to view the data from all available perspectives.

- ***Methodology issues***

  This issue is related to various approaches of data mining and their disadvantages. The examples include diversity of data, versatility of methods, domain dimensions, knowledge assessment, use of domain knowledge, controlling errors in data etc. The best way to resolve this issue is to have a different approach of data mining based on the data being dealt with.

- ***Performance issues***

  It is well known that data mining handles a large amount of data and this is done through the application of statistics and artificial intelligence heuristics. The size of the data is ever increasing and the demand for scalability and efficiency is also growing. The normal algorithm used in data mining is linear algorithms rather than exponential algorithms. Parallel programming is another issue in data mining. The introduction of parallelism will improve the efficiency of data mining methods.

- ***Data source issues***

  Issues related to data can pertain to diversity of data types. The current method of data mining is to collect maximum data as possible regardless of whether the data is taken in the right amount and also if the data taken is appropriate. Data of different types are stored in different repositories. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

VIII. CONCLUSION

Data mining focuses on the extraction and analysis of data from a large amount of data in a database. In this review the basic concepts, applications, challenges and uses of data mining are discussed. This review would be helpful for researchers to focus on challenges and their solutions to data mining.

New constraints and algorithms for enhanced security and more precise data retrieval could be introduced which will help data mining to establish itself as a separate discipline where more studies could be conducted and more improvements made. The future of data mining can extend to genetics through genetic algorithms, artificial neural networks that resemble biological neural networks, automated prediction of trends and behaviors. The extraction of useful if-then rules from data which is called as rules induction is another future trend in data mining. The future cannot be foretold but there are plenty of challenges awaiting solutions and data mining will definitely be a breakthrough in the future of technology and mankind.

REFERENCES

[1] Gorunescu, F, *Data Mining: Concepts, Models, and Techniques*, Springer, 2011.

[2] Han, J., and Kamber, M., *Data mining: Concepts and techniques*, Morgan-Kaufman Series of Data Management Systems San Diego:Academic Press, 2001.

[3] NeelamadhabPadhy, Dr.Pragnyaban Mishra and RasmitaPanigrahi, "*The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)",* vol.2, no.3, June

[4] Heikki, Mannila,*Data mining: machine learning, statistics and databases*, IEEE, 1996.

[5] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P, *From Data Mining To Knowledge Discovery in Databases*", The MIT Press, ISBN 0–26256097–6, Fayap, 1996.

[6] Piatetsky-Shapiro, Gregory,*The Data-Mining Industry Coming of Age,"IEEE Intelligent Systems*, 2000.

[7] Jing He *Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application*, 978-0-7695- 3859-4, IEEE, 2009.

[8] Berry, M.andLinoff, G., *Master Data Mining: The Art and Science of Customer Relationship Management*, Wiley publisher, 2000.

[9] http://maaw.info/DataMining.htm

[10]Jiawei Han, *Data mining concepts and techniques*,2006

[11] Wang, Guoyin, *Granular Computing*,2008

[12]PančePanov, Larisa Soldatova, SašoDžeroski, *Ontology of core data mining* entities, July 2014.

[13] Timothy D Rey, Justin Kauhl, *Using data mining for forecastingproblems*, 2013

[14] S M Sadatrasoul, M R Gholamian, M Siami, Z Hajimohammadai, *Credit scoring in banks and financial institutions via data mining techniques*, 2013

[15] Cesar Astudillo, Matthew Bardeen, NarcisoCerpa, *Data mining in electroniccommerce*, January 2014

[16] Madhumathi K, Dr Antony SelvadossThanamani, *Image mining frameworks and techniques*, 2014

[17] Arun George Eapen, *Application of data mining in medical applications*, University of Waterloo, 2004
[18] AzharMahmood, Ke Shi, ShadeenKhatoon and Mi Xiao, *Data mining techniques for wireless sensor networks*, 2013
[19] IBM, *Using data mining to detect insurance fraud*, 2011
[20] H Lookmansithic, T Balasubramanian, *Survey of insurance fraud detection using data mining techniques*, 2013
[21] UmutAvci, Andrea Passerni, *Improving activity recognition by segmental pattern mining*, 2014
[22] PijushSamui, *"Applicability of Data Mining Techniques for Predicting Electrical Resistivity of Soils Based on Thermal Resistivity." Int. J. Geomech*, 2013
[23] BhavaniThuraisingam, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen, *Data mining for security applications*, 2008
[24] PritamGundecha, Huan Liu, *Mining social media*, 2012
[25]Hengqing Tong, Application of data mining in psychological evaluation,*Computer Science and Computational Technology, 2008.ISCSCT '08. International Symposium*, 2008

Brief Author biography

Sankar K V received the B Tech Degree in Information Technology from Hindusthan College Of Engineering And Technology Coimbatore affiliated to Anna University in 2008 and is currently pursuing M E degree at Hindusthan Institute of Technology Coimbatore affiliated to Anna University.

Dr S.Uma is Professor and Head of PG Department of Computer Science and Engineering at Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India. She received her B.E., degree in Computer Science and Engineering in First Class with Distinction from PSG College of technology in 1991 and the M.S., degree from Anna University, Chennai, Tamilnadu, India. She received her Ph.D., in Computer Science and Engineering Anna University, Chennai, Tamilnadu, India with High Commendation. She has nearly 24 years of academic experience. She has organized many National Level events like seminars, workshops and conferences. She has published many research papers in National and International Conferences and Journals. She is a potential reviewer of International Journals and life member of ISTE professional body. Her research interests are pattern recognition and analysis of nonlinear time series data.