A Review of Data Mining Literature

Article · November 2016

CITATIONS
18

READS
12,904

3 authors:

Sonu Mittal
Jaipur National University
29 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Mirza Shuja
Lovely Professional University
14 PUBLICATIONS **131** CITATIONS
SEE PROFILE

Majid Zaman
University of Kashmir
121 PUBLICATIONS **857** CITATIONS
SEE PROFILE

A Review of Data Mining Literature.

Shuja Mirza
PhD Scholar
School of Computer and Systems Sciences
Jaipur National University, Rajasthan India
Info.shuja@yahoo.in

Dr. Sonu Mittal
Assistant Professor
School of Computer and Systems Sciences
Jaipur National University, Rajasthan India
Sonum7772@rediffmail.com

Dr. Majid Zaman
Scientist B
Directorate of IT&SS
University of Kashmir
Srinagar India
zamanmajid@gmail.com

Abstract **- With progression in technology specifically in last three decades or so, an enormous magnitude of information has been transitioned into a digital form, which resulted in formation of enormous data repositories. With accrual of information in these repositories a challenge persisted as how to extract meaningful knowledge from it. Data mining as a tool was used to tackle the situation. Data mining considered as stepping stone to procedure of knowledge discovery in databases, this is a procedure of extracting hidden information from enormous sets of databases to excavate eloquent patterns and rules. Data mining has now become an indispensable component in almost every field of human life. The present article provides an analysis of the available literature on data mining. The concept of data mining as well as its various methodologies are summarized. Some applications, tasks and issues related to it have also been illustrated.**

Keywords- Data mining, Knowledge discovery in database, Knowledge base.

1. INTRODUCTION

The readiness of ample magnitude of data in almost every field and the desire to excerpt beneficial information and knowledge from it substantiated as main motivation that pulled the eyes of researchers in recent past towards data mining. The information and knowledge extracted can be momentously useful for the applications ranging from small business management to complex engineering design to science exploration. Data mining is the analysis and scrutiny of mammoth data sets, with an aim to uncover significant pattern and rules that were previously unidentified. The core aim is exploiting the data processing power of computer with human's capability to perceive patterns (Han and Kamber 2001)[1]. The epoch of data mining applications was conceived in the year 1980 predominantly by research driven tools engrossed on solo chore (Piatetsky-Shapiro 2000([2]. In recent times data mining is being dominant among Statisticians, MIS communities' data analysts. It was during the first workshop on KDD in 1989 Piatetsky-Shapiro coined the phrase "knowledge discovery in database". The recognition of data mining and KDD shouldn't be astonishing, considering the scale of data been collected from various obtainable sources, the collected data is magnanimous to be examined manually and many a times automatic data analysis supported by classic statistics and machine learning could face concerns once the procedure is hefty and collected knowledge comprises of problematical entities. The bellicose, massive volume of data collected from numerous sources and kept in vast and various repositories. The data collection thus exceeds the human aptitude for analysis without a powerful analysis tool, as a consequence these repositories become 'data vaults', that are not often visited. As decision makers lack tools to extract the treasurable knowledge mounted within enormous volume of data, hence vital decisions lack the utilization of information rich data. Data mining tools perform analysis of data and determine the vital patterns that were earlier anonymous. As every arena of human life has become data intensive which stemmed in making data mining as an indispensable constituent. Though Data mining and KDD have been used conversely yet KDD can be seen as an inclusive procedure of extracting beneficial knowledge from data, while as Data mining can be seen as core of KDD, which includes Algorithms that explore data, build models and discovery unknown patterns.

2. REVIEW OF LITERATURE

Fayyad et.al (1996)[3] in their paper "From data mining to knowledge discovery in databases" described KDD as "a nontrivial process of recognizing valid, novel, potentially useful and finally understandable patterns in data". Elaborating the definition data were any set of valid facts that are accessible in an electronic form. Patterns are models expressed by some language as data subset. The patterns must be valid so that are true and can be modeled for any new data. Process includes multiple steps from data preparation to knowledge enhancement all recurrently used till looked-for results are achieved. Nontrivial indicates that there ought to be a sort of inference computation so as to distinguished it from the traditional computation of values. Fayyad and Stolorz (1997)[4] in their paper described KDD as "generalized procedure of uncovering treasured knowledge from data with mining being one among other steps in that process that uses some algorithms for knowledge extraction process". Charles et.al (1998)[5]

proposed data mining as an effective tool for direct marketing so as to improve product marketing in this technological age where traditional means of marketing such as mass marketing is showing downfall trend. Using data mining we can determine buyers' patterns, in order to single out potential buyers from customers list. It was demonstrated that data mining as direct marketing tool can bring more profit than the traditional means of mass marketing as it targets only the potential buyers. Michael Goebel et.al (1999)[6] in their paper "A survey of data mining and knowledge discovery tools", provided an generalized view of common knowledge discovery tasks and various methodologies to resolve these. A feature classification scheme was proposed that was used to study knowledge and data mining software's. They specified some of the important features that must be reckoned important for the knowledge discovery software so that it will be used effectively and will address more issues that were not sufficiently studied. lots of the organizations in the world today have very huge databases that don't have any limitation on growth. New data is added to these databases at rate of millions of records per day. These types of databases provide a new challenge and unique opportunities to mine these data streams. Pedro Domingos et.al (2000)[7] described and evaluated VDTF on these huge databases. They used Hoeffding trees which allow learning in a meager constant time per example and it has high asymptotic similarity to batch time. Differentiating data mining from information Hand et.al (2001)[8] defined data mining as "analysis of huge datasets to discover the unsuspected relationship and to review data in more logical way so that it serves the desired results". According to Rygielski et.al (2002)[9], data mining technology has added a new dimension to CRM. The data mining's' power to extract the predictive unknown information from vast datasets have found its way into the CRM to identify and evaluate valuable customers, predict the customers shopping behavior which results in helping the vendors taking proactive and knowledge based decisions. Eamonn Keogh et.al (2004)[10], discussed "parameter free data mining" as parameter laden algorithms may over or under estimate certain parameters which will yield in patterns that may not be fully accurate. A parameter free mining can prevent us from applying our own presumptions or prejudices. They proposed a datamining paradigm centered on compression. Streaming datamining considered as hard chore in knowledge discovery in databases, the traditional mining approaches are infeasible to deal with it as data comes in multiple, unceasing and time wavering data streams. Alhammady et.al (2007)[11], introduced a unusual methodology for mining evolving patterns in streaming data which has a better mining complexity and classification accuracy as proved by experimentation. As most of current methodologies of data mining explore knowledge in single data table. But recently most of these methodologies are protracted to relational cases. Relational data mining involves applying data mining approach on multiple table data for abstracting the knowledge in it (Saso Dzeroski 2010)[12]. Venkatadri.M et.al (2011)[13], discuss appropriate Techniques and methodologies are needed in future to cater the needs of data mining field as it is exploring more and more complex fields so that we can explore the such complex situations where data is huge but is full of hidden information. Tipawan Silwattananusarn et.al (2012)[14] in their review paper reconnoitered the applications of data mining techniques that evolved overtime to support knowledge management process as it is being extensively used across various fields and each field is being supported by discrete data mining techniques, it was shown that data mining can be integrated into knowledge management framework and improve that process with superior knowledge. Divya Tomar et.al (2013)[15] presented data mining as most vivacious and appealing research area which is gaining attractiveness in medical domain. Data mining provides several benefits in healthcare domain. It enhances the medical services in cost effective manner. Anand.v. saurkar et.al (2014)[16] defined data mining as "interdisciplinary field which consists of integrated databases, artificial intelligence, machine learning, statistics etc.". They defined data mining as multi-step process which comprises preparation of data for mining, mining algorithms, analysis of results and interpretation of results. The capability of data mining to dig deep into the data and extract hidden information and knowledge from it has received tremendous attention form business professionals to generate the patterns related to customer's behavior and predict future sales and trends, and assist policy makers in decision making with the aim of increasing profits (Shraddha Soni 2015)[17].

3. ARCHITECTURE OF DATA MINING

The architecture of data mining is shown in figure 1

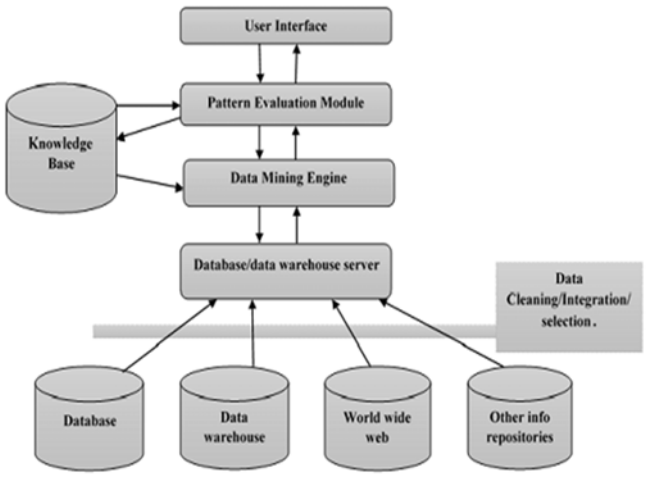

Figure 1. Architecture of Data Mining. (credits: Han and Kamber [1]).

Knowledge Base: It serves as the initiation for the whole data mining process. It acts as guide for searching or assessing the interestingness of the resultant patterns. Such type of knowledge may include concept hierarchies, that organize attributes or their values into distinct stages of abstraction.

Data mining engine: It forms the staple component of mining system, it consists of all the necessary modules such as characterization, prediction, cluster analysis, outlier analysis and evolution analysis for performing data mining tasks.

Pattern evaluation module: This module is usually associated with interestingness measures. It persistently interrelates with the data mining engine to remain focused on search of interesting patterns. Many of a times it uses thresholds to sieve out discovered pattern, or may use pattern evaluation module integrated with mining module, depending on data mining technique used.

User interface: The module acts as a connection between users and data mining system. It facilitates users' interaction with the system in an easy and efficient way without fretting the user about convolutions behind the process.

Data sources (www, data warehouse, database, other repositories): These are the actual sources of data and enormous volume of historical data is required for successful data mining. Organizations typically store data in databases or data ware houses. Sometimes more than one databases or text files or spreadsheets are contained in data warehouse. www is an another huge source of data.

Database or data warehouse server: It contains tangible data that is set to be fetched. Fetching of data on users' request is its key responsibility.

Other process: before data is passed on into the data warehouse server, data needs to be cleaned and integrated, as data is collected from distinct sources and is in different formats so it can't be used directly for mining process. The data needs to be cleaned, integrated and only the reliable data needs to be selected and passed on to the data warehouse server. The process may require number of techniques for cleaning, integration and selection.

4. DATA MINING PROCESS

Fayyad et.al (1996)[18] defined "data mining as one among several steps in the process of knowledge discovery, it involves applying data analysis and discovering algorithms that yield a precise enumeration of pattern over data under any acceptable computation efficiency". This procedure is collaborative and reiterative and encompasses of many steps with decisions made by user with attempts made at every step to complete a particular discovery task, each accomplished by the application of discovery method. Data mining synonymously used by some for phrase knowledge discovery in databases (KDD) process, conversely many consider it as an essential step of KDD which results in beneficial patterns or models for data. The various processes for data mining are shown in figure 2.

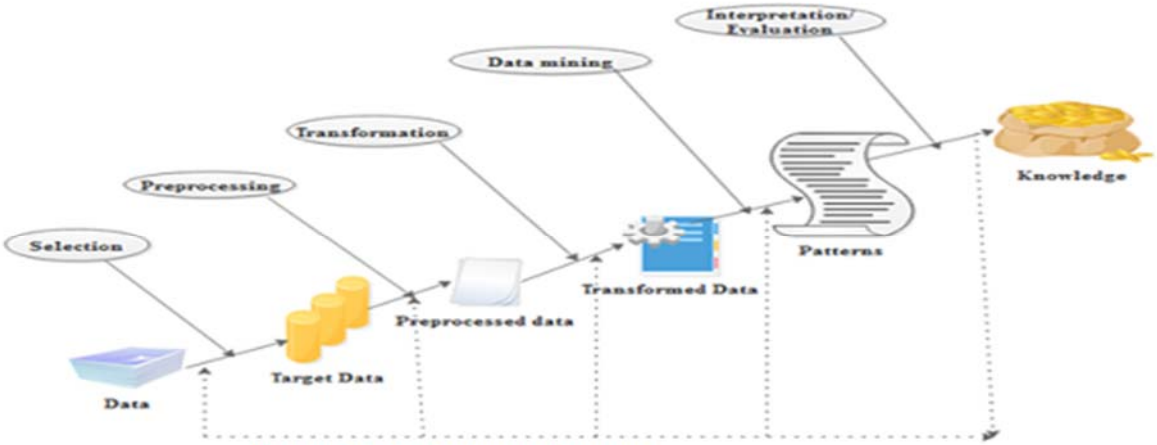

Figure 2. Data Mining Process

Selection: Selecting pertinent data from distinct sources for mining process.

Preprocessing: As data is congregated form different sources, it contains inconsistencies, to remove those various activities are carried out in this phase, blemished data is corrected or removed, noise and

discrepancies are removed and data from distinct sources is combined.

Transformation: Here data is transmuted into appropriate form for mining. Feature selection, sampling, aggregation may be used.

Data Mining: it is a significant step here mining algorithm is chosen which is apposite to pattern in data. Extraction of data patterns is also carried out.

Interpretation and Evaluation: To recognize and infer mining results or patterns into knowledge by elimination of redundancies and irrelevant patterns. Here an assortment of visualization and GUI stratagems are used for transforming the advantageous patterns into the human understandable terms.

5. DATA MINING TASKS

Data mining tasks are grouped into two main categories 1: Predictive 2: Descriptive. These two are considered primary objectives of data mining. Fayyad et.al 1996 define six main functions of data mining: 1. Classification 2. Regression 3. Clustering 4. Dependency modelling 5. Deviation detection. 6.Summarization.

Classification, regression and anomaly detection categorized under predictive category while as clustering, Dependency modelling categorized under descriptive category. Predictive model forecasts using some variable in dataset so as to predict unknown values of other relevant variable while as descriptive model classifies patterns or relationship and encompasses human understandable pattern and trends in data (Gorunescu Florin 2011)[19].

Classification: classification is among the classical data mining technique that is established on machine learning. It finds mutual properties amongst a set of objects in a database and categorizes them into diverse classes in accordance with the classification model. Its main objective is to scrutinize the training data and develop an accurate description or model for each class using feature available in data. This method uses mathematical techniques like decision trees, Neural networks and statistics (Ming-Syan et.al 1996)[20].

Regression: It is one among data mining techniques that defines the association between dependent and independent variables. Prediction is accomplished with regressions support. Statistically regression is the mathematical model that constitutes connection amongst the values of dependent variable and values of other predictor or independent variable. In regression the predicted variable may be continuous variable. In regression real valued prediction variables are mapped from items of a learning function. Statistical regression, Neural Network, Support Vector Machine regression are some of the commonly used regression strategies. More complex techniques such as Logistic regression, Decision Trees or Neural Networks could also be utilized for forecasting future values, these techniques could also be combined for attainment of better result.

Clustering: it is a data mining technique which groups physical or abstract objects into classes of similar objects [19]. clustering is a method of dividing set of data(records/tuples/objects/samples) into several groups(clusters) based on foreordain similarities. The principal aim of clustering is finding groups(clusters) of objects based on affinity so that within individual cluster there is great resemblance to each other while clusters are diverse enough from one another. In machine learning terminology, clustering is a form of unsupervised learning [19].

Dependency Modelling (Association Rule Mining): it's amongst the finest acknowledged data mining techniques and is categorized under unsupervised data mining technique, which aims at finding connections or relations between items or records belonging to a large dataset and labels significant dependencies among variables. Association rule mining is implication of the form $X \rightarrow Y$, where x and y are distinct items or item sets manufacturing if-then statements regarding attribute values. In market basket analysis this rule has been commonly used, it tries to analyze customers purchasing certain items and provides insight into the combinations customer frequently purchases together.

Anomaly detection: synonymous to is its name it deals with the unearthing of most substantial changes or aberrations from the standard behavior [19].

Summarization: Though not amongst the techniques of data mining, but is a resultant of these techniques and deals with determining a compact depiction for a subset of data synonymously referred to as generalization or description.

Sequential Patterns: Sequence discovery is a data mining technique that is used to determine sequential patterns or associations or regular events/trends between variable data fields over a business period.

6. ISSUES IN DATA MINING

With data mining being well developed but it still faces variety of issues with its practical implementation [1], some of them discussed below:

Security issues: Security is the most critical and vital issue concerning any data transaction process, given extremely confidential nature of data, potential illegal access to the knowledge should be prevented and secrecy must be guarded.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Mining Methodology issues: as different users have interest in diverse kinds of knowledge, data mining must cover a wide spectrum of data analysis and knowledge discovery tasks which may use same data base in different ways and require development of numerous data mining techniques.

User interface issues: Knowledge discovery by data mining tools is advantageous and expressive only as long as it is presented explicitly and fascinatingly to the used. As it is difficult to know what can be discovered within database, the mining process should be interactive, data should be presented in high level language, visual representations or other graphical expression forms, so that user can understand and interpret it and use it as required.

Handling noisy and incomplete data: Data stored in databases can be varying as various issues are associated with data sources, the data may be incomplete, the data may contain cases which may raise exceptions. Mining data with these irregularities raise ambiguities in the process, causing knowledge model constructed to over-fit data and truncate the accuracy of resultant knowledge, so mining methods that can handle these inconsistencies are required.

Performance issues: For datamining efficiency and scalability are strategic factors for datamining implementing a database system. Information must be effectually and proficiently extracted from databases as they are enormous in amount. The algorithms used should be proficient and scalable, their running time needs be predictable and acceptable for large databases.

7. APPLICATIONS OF DATA MINING

Application of Data Mining in Health Care: Data Mining can be meaningfully advantageous in healthcare system, but its success axles on availability of clean data. In healthcare it is used for diagnosis and prognosis of disease, also affiliations among disease can be established. Physicians can identify effective and best practices so that patient gets improved and reasonable services. As enormous amount of healthcare data is complex and vast to be processed and analyzed, datamining provides methodology and tools for transformation of data into information for efficacious decision making (Parvathi I .et.al. 2014)[21].

Application of Data Mining in Educational Systems: Data mining in educational system is an evolving field with researchers showing deep interestingness towards it. As millions of students are enrolled each year in different institutions thus adding immense volume of data. Data Mining techniques can help in spanning the knowledge gap in educational system by ascertaining veiled patterns, connotations and variances. This helps stakeholders to improve efficacious decision making which results in refining educational system (C.Romero et.al 2005)[22].

Application of Data Mining in CRM: Data mining in CRM is currently most discussed topic of research in industry and academia with the aim of giving research summary on utilization of data mining techniques in CRM domain (EWT Ngai et.al 2009)[23].

Application of Data Mining in Market Basket Analysis(MBA): Various techniques of Data Mining are used for market basket analysis-MBA, a technique which helps in finding out the association between various items which a customer puts in his shopping cart during shopping, it observes shopping habits of customers. The business houses can use datamining techniques to identify the buying patterns and behavior of customer based on which a range of choices can be presented to customer as per his habit of buying (T Raeder 2011)[24].

Application of Data Mining in Sports Data: Data mining techniques have also infringed into the field of sports. A huge number of games are being played with each sport generating massive amount of statistical data. This massive data needs to be maintained with regarding to scheduling of events and statistics of individual player in these events. Data mining can be used for forecasting and analysis of performance and for strategy planning (O.K Solieman 2006)[25].

8. CONCLUSION

The paper presented a revision of literature vis-à-vis data mining, a technique used to ascertain hidden and useful patterns from vast amount of datasets. These discovered trends help originations to predict the future behavior of customers or products. This study gives the idea about various data mining techniques, different methods, different processes and some issues related to datamining. In future we tend to review and compare various algorithm's used in datamining.

9. REFERENCES

[1]. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[2]. Piatetsky-Shapiro, Gregory. "Knowledge discovery in databases: 10 years after." ACM SIGKDD Explorations Newsletter 1.2 (2000): 59-61.

[3]. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.

[4]. Fayyad, Usama, and Paul Stolorz. "Data mining and KDD: Promise and challenges." Future generation computer systems 13.2 (1997): 99-115.

[5]. Ling, Charles X., and Chenghui Li. "Data Mining for Direct Marketing: Problems and Solutions." KDD. Vol. 98. 1998.

[6]. Goebel, Michael, and Le Gruenwald. "A survey of data mining and knowledge discovery software tools." ACM SIGKDD explorations newsletter 1.1 (1999): 20-33.

[7]. Domingos, Pedro, and Geoff Hulten. "Mining high-speed data streams." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[8]. Hand, David J., Heikki Mannila, and Padhraic Smyth. Principles of data mining. MIT press, 2001.

[9]. Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. "Data mining techniques for customer relationship management." Technology in society 24.4 (2002): 483-502.

[10]. Keogh, Eamonn, Stefano Lonardi, and Chotirat Ann Ratanamahatana. "Towards parameter-free data mining." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[11]. Alhammady, Hamad. "A novel approach for mining emerging patterns in data streams." Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on. IEEE, 2007.

[12]. Džeroski, Sašo. Relational data mining. Springer US, 2009.

[13]. Venkatadri, M., and Lokanatha C. Reddy. "A review on data mining from past to the future." International Journal of Computer Applications 15.7 (2011): 19-22.

[14]. Silwattananusarn, Tipawan, and Kulthida Tuamsuk. "Data mining and its applications for knowledge management: a literature review from 2007 to 2012." arXiv preprint arXiv:1210.2872 (2012).

[15]. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[16]. Saurkar, Anand V., et al. "A Review Paper on Various Data Mining Techniques." International Journal of Advanced Research in Computer Science and Software Engineering 4.4 (2014).

[17]. Soni, Shraddha. "A Literature Review on Data Mining and its Techniques." Indian Journal of Applied Research 5.6 (2016).

[18]. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.

[19]. Gorunescu, Florin. Data Mining: Concepts, models and techniques. Vol. 12. Springer Science & Business Media, 2011.

[20]. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and data Engineering 8.6 (1996): 866-883.

[21]. Parvathi, I., and Siddharth Rautaray. "Survey on data mining techniques for the diagnosis of diseases in medical domain." International Journal of Computer Science and Information Technologies 5.1 (2014): 838-846.

[22]. Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146.

[23]. Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36.2 (2009): 2592-2602.

[24]. Raeder, Troy, and Nitesh V. Chawla. "Market basket analysis with networks." Social network analysis and mining 1.2 (2011): 97-113.

[25]. Solieman, Osama K. "Data mining in sports: A research overview." Dept. of Management Information Systems (2006).