CHALLENGING ISSUES OF WEBMINING TECHNIQUES

R TEJASREE,Y SAMYUKTHA REDDY,S MEGHNA,B NIHARIKA, K SWATHI
STUDENTS,
Department of CSE,
Megha Institute of Engineering and Technology, Edulabad,Ghatkesar(Md),Meddchal(Dt),TS,India

Abstract–

Data mining on large databases has been a major concern in research community, due to the difficulty of analyzing huge volume of data using only traditional OLAP tools(Online Application Processing). This sort of process implies a lot of computational power , memory and disk input or output , which can only be provided by parallel computers . This research paper also conducts a formed review of application of data mining such as Education , Banking , Insurance , Medicine , Manufacturing Engineering , Health care , Transportation , Research analysis , Sales and Marketing . This paper provides a survey of various data mining techniques .These techniques are Classification analysis , Association rule learning , Anomaly or outlier detection , Clustering analysis and Regression analysis . This paper discuss the topic based on past research paper and also studies the data mining techniques and application.

INTRODUCTION

In the real world , huge amount of data are available in education , medical , industry and in many other areas .

Data Miningis a type of sorting technique whichis actually used today by companies with astrongconsumer focus retail , financial,communicationand marketing organizations . It enables these companies to determine relationshipsamong“Internal” factors such asprice ,product positioning , orstaff skills , and “External” factors such as economic indicators,competitions, and customerdemographics.

Forexample ,blockbluster entertainment minesits video rental history database to recommend rentals to individual customers. American expresscan suggest products to its card holdersbased on analysis of their monthly expenditures.

Walmart is pioneering massivedata mining to transform its supplier relationships . Walmart allows more than3,500 suppliers, toaccessdata ontheir products and perform dataanalysis.These suppliers use this datatoidentify customer buying

patterns at the store display level. In 1995,Walmart computers processed over 1millon complex data queries.

The National Basketball Association (NBA) is exploring a data mining application that can be used in conjuction with image recordings of basketball games. The advanced scout software analyzes the movements of players to help coaches orchestrateplays and strategies.

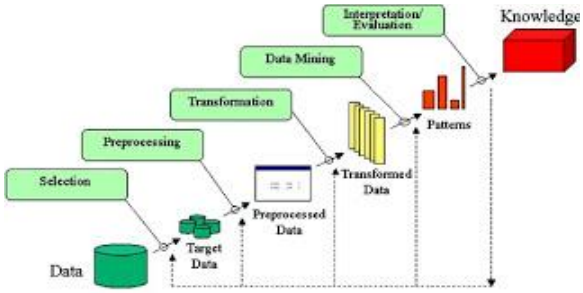

Fig: data mining database.

Data mining software analyzes relationships and patterns in stored transactions data based on open-ended user queries . Generally anyfour types of relationships are sought :

Classes : Stored data is used to locate data in predetermined groups .

Clusters : Data items are grouped according to logical relationship or consumer preferences .

Associations : Data can be mined to identify associations .

Sequential patterns : Data is mined to anticipatebehavior patterns and trends .

Process

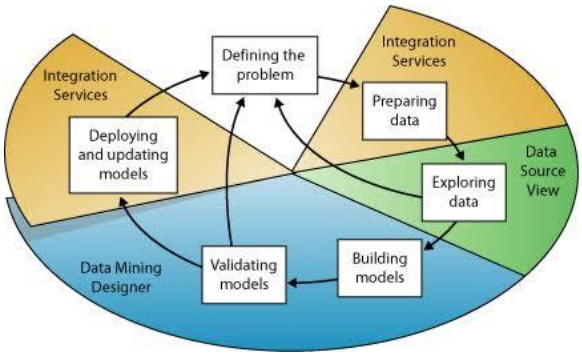

Fig: process for data mining.

DATA MINING APPLICATIONS

Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data .Data mining applications areas includes marketing, telecommunications, finance , education , medical and so on . Some of the main applications are listed below:

1.Data mining in education sector : We are applying data mining in education sector then new emerging field called “EDUCATION DATA MINING” . The goals of EDM are identified as predicating students future learning behavior ,studing the effects of educational support , and advancing scientific knowledge about learning . Data mining can be used by an institution to take accurate decisions and also to predict the result of the students .

2.Data mining in banking and finance: Credit and spending by customer groups can be identified by using data mining . the hidden correlations between different financial indicators can be discovered by using data mining . Data mining isused toidentify customersloyality byanalyzing thedata of customers purchasingactivities . In the financial markets , datamining techniquesuch as neural

networksused in stock forecasting , price predicationand so on .

3.Data mining in insurance: Data mining is applied in claims analysis such as identifying which medical proceduresare claimed together . Data mining enables toforecasts whichcustomerswill potentially purchase new policies . Ithelps todetect fraudulent behavior .Italsoallows insurance companies to detect riskycustomersbehavior patterns .

4.Data mining in medicine: Data mining enables to characterize patientactivities to see incomingoffice visits .It helpsin identifying the patterns of successful medical therapies for different illnesses.Example : smarthealth prediction in data mining .

5.Data miningin manufacturing engineering: Data mining tools can bevery useful to discover patternsin complex manufacturing process. Data mining can be used in system-leveldesigningto extract therelationships betweenproduct architecture, product portfolio, and customer needs data .It can also be usedto predictthe product development span time, cost ,and dependenciesamong other tasks.

6. Data mining in health care: Datamining applications can be developed to evaluate the effectiveness of medicaltreatments . By comparing and contrasting causes ,symptoms and courses of treatments ,data mining can deliverananalysis of which courses of actionsprove effective. In 1999,florida hospital launched the clinical best practices initiatives with the goal of developing a standard path of care across all campuses , clinicians and patient admissions .

7.Data mining in transportations: Data mining helps determine the disturbution schedules among ware houses and outlets and analyses loading pattrerns.

8. Data mining in research analysis: History shows that we have witnessed revolutionary changes in research . Data mining is helpful in data cleaning , data pre-processing and integration of databases .The researchers can find any similar data from the data base that might bring any change in the research . Data visualisation and visual data mining provides us with a clear view of data .

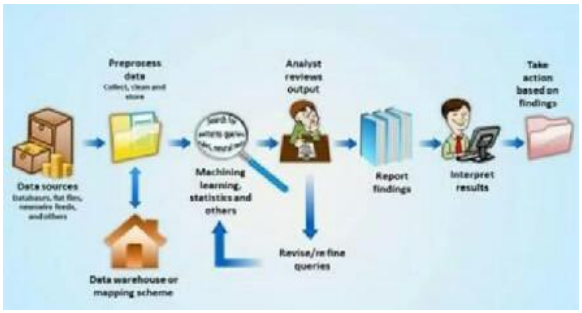

Fig : data mining applications.

9.Data mining in sales and marketing: Data mining is used for “Market Basket Analysis” to provide information on what product combinations were published together when they were bought and in what sequence this information helps business promote their most profitable products and maximize the profit in addition to it encourages the customers to purchase related product that they may have been missed or overlooked .

TECHNIQUES

There are five types of techniques .They are

1.Classification Analysis: Thisanalysis is used to retrieve important and relevant information aboutdata , and metadata . It is used to classify different data in different classes . Classifications is similar to clustering ina way thatit also segments data records intodifferent segments called classes. But unlike clustering , here the data analysts would have acknowledge of different classes or

cluster .So ,in classifications analysis you would apply algorithms to decide how new data should be classified . A classicexample of classification analysis would beour outlook email.

2.Association RuleLearning : It refers to the method that can help you identify some interesting relations between different variables in largedatabases . This technique can help you unpacksome hidden patterns in the data that can be used to identify variables within the dataand the concurrence of different variables that appear very frequently in the dataset . Association rules are useful for examining and forecasting customer behavior.It ishighly recommendedin the retail industry analysis . In IT ,programmers use association rules to bulidprograms capable of machine learning .

3.Anomaly or Outlier detection: This refers to the observation for data items in a dataset that do not match an expected pattern or anexpected behavior . Anomalies are also known as outliers, novelties , noise,deviations,and exceptions . Often theyprovide critical and actionable information . These types of items are statistically aloof as compared to the rest of the data and hence , it indicates thatsomething out of the ordinary has happened and requires additional attention . This technique can be used in a variety of domains ,such as intrusion detection , system health monitoring , fraud detection ,fault detection ,event detection in sensor networks ,and detecting eco-system disturbances . Analysts often remove the anomalousdata from the dataset top discover results with an increasedaccuracy .

4. Clustering analysis: The cluster is actually a collection of dataobjects ; thoseobjectsare similar within the same cluster . That means theobjects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters . Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise . A result of this analysis can be used tocreate customer profiling .

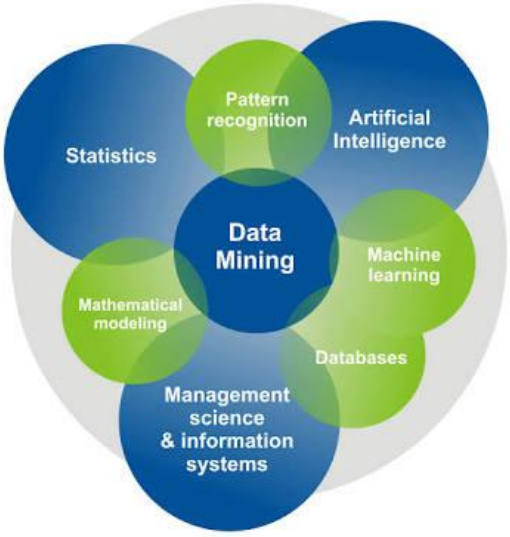

fig: techniques involved in data mining .

5.Regression analysis: In statistical terms ,a regression analysis is the process of identifying and analyzing the relationship among variables . It can help you understand the characteristicvalue ofthe changes , if any one of the independent variableis varied . Thismeans one variable is dependent on another ,but itis not vice versa . It is generally used for predication and forecasting .

Allthese techniques can help analyze different data from different perspectives .

FUTURE DIRECTIONS

First , themostfocused and extensively studied topic in frequent pattern mining is perhaps scalable mining methods . When we are working with data streams still it is aresearch challenge to derive a compact but high quality set of

patterns that are most useful in applications .The set of frequent patterns derived by most of the current patterns mining methods including ours give approximate patterns as stream is flowing continuously and some data is lost in the process of analyzing the stream .

To make frequent patterns mining an essential task in data mining , much research is needed to further develop pattern-based mining methods . For example ,classification is an essentialtask in data mining . Construction of better classificationmodels using frequent patterns than most other classification methods is again a research issue .

Another major research area in frequent mining is interpretation ofpatterns i.e. , semantic annotation for frequent patterns ,and contextual analysis of frequent patterns . The semantics of a frequent patterns includes deeper information . What is the meaning of thepatterns ; What are the synonym patterns ;andWhat arethe typical transactions that this patterns resides ?

On one side , it is important to go to the core part ofpatterns mining algorithms, And analyze the theoretical properties of differentsolutions. Much work is needed to explore new applications of frequent patterns mining . For example , bioinformatics has raised a lotof challenging problems , and we believe frequent pattern mining may contribute a good dealto it with further research efforts.

Achievements: In this these our objective was to :

- Construct synopsis of data stream of transactions .
- Mine frequent itemsets.
- Mine frequent patterns.
- Mine infrequentpatterns .

Constructsynopsis of data stream oftransactions : Thedifferent techniques related tosynopsisconstruction with special emphasis on reservoir sampling . We have proposed two algorithms based on reservoirsampling to construct synopsis and to mine frequentitemsets .

Mine frequentitemsets: We have proposed a new counter based algorithm to mine frequent itemsets . This work is published in an internationaljournal .

Mine frequent patterns : In this we proposed a new data structure called Dynamic FP-treetomine frequent patterns . Experiments havethe efficiency of dynamic fp-tree.

Mine infrequent patterns: we have proposed a new algorithm based on dynamic FP-tree to mine infrequent patterns. Thisworkispublished in an internationaljournal .

The silent features of the research work carried out are:

- A through literature review has been carried out .
- Data stream iscontinuous flow of data . Since it is not possible to analyzethe wholestream . Hence a new algorithm based on reservoir sampling has been proposed to construct synopsis of datastream .
- Frequentitemset mininghas number of scientific and commercial applications . This algorithm is also based on reservoir sampling.

Conclusion

Frequent pattern mining has been a focused themein data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made

, ranging from efficient and scalable algorithms for frequentitemset mining in transaction databases to numerous research frontiers , such as sequential pattern mining,structured pattern mining , correlation mining , associative classification ,and frequent pattern-based clustering , as well as their broad applications . It is believed that frequentpattern mining research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run . However , there are still some challenging research issues that need to be solved before frequent pattern mining can claim a cornerstone approach in data mining applications .

REFERENCES

1. Hand D., Manila H . and Smyth P., Principles of Data Mining , MIT Press ,2001.
2. Achlioptas , P., Scholkopf , B., and Borgwardt , K.(2011). In ACM SIGKDD Conference on data mining .
3. Azencott , C., Grimm, D., Sugiyama , M., Kawahara , Y., and Borgwardt , K. M. (2013) .Bioinformatics .
4. Bishop ,C. M.(2006) . Pattern Recognition