ISSN (Online): 2455-7838
SJIF Impact Factor (2016): 4.144

EPRA International Journal of

Research & Development

(IJRD)

Monthly Peer Reviewed & Indexed International Online Journal

Volume:2, Issue:8, August 2017

Published By :
EPRA Journals

ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
SJIF Impact Factor: 4.144 Volume: 2 | Issue: 8 | August | 2017

ANALYSIS OF UTILITY BASED PATTERN MINING USING NEURAL NETWORKS

Sujamol S[1]
[1]Assistant Professor, Department of Computer Science & Engineering,
Sree Narayana Guru Institute Science and Technology,
Eranakulam, Kerala, India

Anoob C S[2]
[2]Assistant Professor, Department of Electronics and Communication Engineering,
Sree Narayana Guru Institute Science and Technology,
Thekkethazham, Eranakulam, India

ABSTRACT

Mining is the process of finding a small set of valuable nuggets from a great deal of raw materials. Text mining is the discovery of interesting knowledge in Text documents. Almost all enterprise collects a variety of information. Because of the explosive growth in the number of mobile phones, bank data, research project and a number of government agencies, it is a challenging issue to find accurate and valuable knowledge and information from mountains of accumulated data. The field of data mining provides some techniques and tools for handling this large amount of information. The interesting patterns which are found must be understandable and actionable. Existing Text mining methods have some drawbacks as they were term-based and suffer from problem of polysemy and synonymy. Although a number of relevant algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. This paper presents a brief survey on common data mining techniques and application of Neural Networks and Genetic algorithm in this field.

KEYWORDS: *Data Mining, Neural Networks, Genetic Algorithm*

I.INTRODUCTION

As computer networks become the backbones of science and economy, enormous quantities of machine readable documents become available. There are estimates that 85%of business information lives in the form of text. Unfortunately, the usual logic-based programming paradigm has great difficulties in capturing the fuzzy and often ambiguous relations in text documents. Many text mining methods have been developed in order to achieve the goal of retrieving information for users. This paper focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. The process of knowledge discovery may consist as following

•Data Selection
•Data Processing
•Data Transaction
•Pattern Discovery
•Pattern Evaluation.

Text mining is also called as knowledge discovery in databases. Data analyzing in knowledge discovery of databases aims at finding hidden patterns as well as connections in those data. Searching for keywords or phrases in a collection are now widely used. Such search only marginally supports discovery because the user has to decide on the words to look for. On the other hand, text mining results can suggest useful patterns to look at, and the user can accept or reject these patterns according to their interest.. This paper presents a survey on common data mining techniques which effectively finds interesting patterns that are actionable

II THE KNOWLEDGE DISCOVERY PROCESS

There is still some confusion about the terms Knowledge Discovery in Databases (KDD) and data mining. Often these two terms are used interchangeably. The term KDD is used to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is also defined as the extraction of patterns or models from observed data. Although at the core of the knowledge discovery process, this step usually takes only a small part of the overall effort. Hence data mining is just one step in the overall KDD process. Other steps involve:

- Developing an understanding of the application domain and the goals of the data mining process
- Acquiring or selecting a target data set
- Integrating and checking the data set
- Data cleaning, pre-processing, and transformation
- Model development and hypothesis building
- Choosing suitable data mining algorithms
- Result interpretation and visualization
- Result testing and verification
- Using and maintaining the discovered knowledge

In practice the two fundamental goals of data mining tend to be: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest. Description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differ with respect to the underlying application and the technique.

III METHODS OF DATAMINIG

- Association: Association (or relation) is probably the better known and most familiar and straight forward data mining technique. Here, a simple correlation between two or more items is made, often of the same type to identify patterns.
- Classification: Classification is used to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. Additionally, classification is used as a feeder to, or the result of, other techniques.
- Clustering: Clustering technique is useful to identify different information by considering various examples and one can see where the similarities and ranges agree. By examining one or more attributes or classes, it is possible to group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results.
- Prediction: Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. By analyzing past events or instances, prediction about an event can be made.
- Sequential Patterns: Various uses of sequential patterns for identifying trends, or regular occurrences of similar events can be traced. For example, with customer data it is possible to identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, this information is used to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

IV.LITERATURE REVIEW

Text Mining and Knowledge Discovery

Text mining is basically an extraction process where efficient methods are used to find useful patterns. Text mining is a knowledge

discovery process where the usual steps of data mining and statistical procedures were applied. Data analysis in the Knowledge Discovery aims at finding the hidden pattern as well as connections in those data. The concept of Text mining was initially introduced during the 1980s, based on manual techniques. But these manual techniques were labour intensive and so costly. It takes lot of time to manage the growing amount of data. There were increasing success in making programs to mechanically manage the data, and within the last ten years there has been a lot of progress. Currently the study of text mining uses the concept of assorted mathematics, applied mathematics, linguistic and pattern recognition techniques which permit automatic analysis of unstructured information and result in useful knowledge, and forms a text that is highly searchable. Initially, this concept of knowledge discovery from the text (KDT) is introduced in Feldman et al. which deals with the machine supported text information analysis. The method includes the retrieval of information, natural language processing and extraction of information and further connects them with methods and algorithms of KDD, machine learning, statistics and data mining. And hence an identical procedure likes the KDD method, where text documents are focused for the analysis was selected.

V.RELATED WORKS

Many types of text representations have been proposed in the past. A well-known one is the bag of words where keywords (terms) as elements in the vector of the feature space were used. In [3], the tf*idf weighting scheme is used for text representation in Rocchio classifiers.tf*idf measure states that tf is the term frequency,idf is the inverse document frequency. Assign a tf*idf weight to each term in document. In addition to TF*IDF, the global IDF and entropy weighting scheme was proposed which improves performance by an average of 40 percent. Various weighting schemes for the bag of words representation approach were given. The problem of the bag of words approach is to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting is difficult. In order to reduce the number of features, many dimensionality reduction approaches have been conducted which uses the feature selection techniques, such as Information Gain, Mutual Information, Chi-Square ,Odds ratio etc.

In 1988, Gerard Salton and Christopher Buckley proposed a paper named Term Weighting Approaches in Automatic Text Retrieval [1]. This system uses text indexing by appropriately using weighted single terms which can help in information retrieval process. Effective term weighting systems are used which can represent the documents in term vector space. According to the weights of terms, the documents are ranked. But it suffers from the problem of polysemy and synonymy.A term with higher tf*idf will be meaningless in some documents and it is difficult to select a limited number of features among an enormous set of words.

In the paper "Mining Sequential Patterns" [2], two algorithms AprioriSome and AprioriAll were proposed. Here each transaction from database of customer transaction was mined. But they have lower performance when no of transaction increases. Sequential pattern mining algorithm mine the full set of frequent subsequence's satisfying a minimum support threshold in a sequence database. But these long sequences contain a large no of frequent sub sequences and it is expensive in both time and space. So a new technique Clospan for Mining Closed Sequential Patterns in Large Databases was proposed in 2003[4].Here only frequent closed sub sequences are mined instead of mining complete set of frequent subsequences.That is the closed sequential patterns does not contain a super sequences with same support. But it does not incorporate user-specified constraints in mining of closed sequential patterns and it will not prune any noise patterns in the discovered pattern space.

In 2004, Ding-Ying Chiu, Yi-Hung Wu, Arbee L. P. Chen proposed a method where a Frequent Sequences mining algorithm by a New Strategy without Support Counting[5] was devised. Here an efficient algorithm known as Direct Sequence Comparison (DISC) was developed which recognizes all frequent sequences without counting support of non-frequent sequences. It combines candidate generation and support counting together with candidate sequence pruning. But Support values and weights of each term is not calculated.

In 2006 a method for Mining Ontology for Automatically retrieving Web User Information Needs [6]was proposed. The fundamental objective of this research is the automatic meaning discovery other than pattern discovery. It was difficult to obtain the right information from the Web for a particular Web user or a group of users because of the difficulty of automatically acquiring Web user profiles. This paper presents a new approach for this problem. It proposes a method for capturing evolving patterns. In addition, it establishes the process of assessing relevance. This paper provides both theoretical and experimental evaluation for this approach. But the problem is that the users cannot distinguish between relevant and irrelevant data. Web users don't know how to represent the interesting topics and a phrase based approach is used which is having low frequency of occurrence.

Inorder to overcome all the above problems, a paper was proposed in 2012 which find efficient patterns in text documents [7].It developed the concept of a Pattern Taxonomy Model(PTM) which included pattern evolving and pattern deploying. A PTM is implemented in three steps.

1) Discovering useful patterns through sequential closed pattern mining algorithms and employs a pruning scheme.
2) Using the discovered patterns by pattern deploying.
3) Using a shuffling algorithm to adjust term support within the pattern.

Here a Pattern Taxonomy Model was developed where a new pattern based model was used for representing text documents. It is a tree like structure that illustrates the relationship between different patterns extracted from text collections. The root of the tree is one of the longest patterns. Here for simplicity, all documents are split into different paragraphs. So a given document say d, it contain a set of paragraphs PS(d). Assume that D is a training set of documents and it consists of a set of positive documents, denoted by D+; and a set of negative documents, denoted by D-. Let term set is represented as T = {t1; t2; . . . ; tm} and it can be extracted from the set of positive documents, D+.Patterns that occur frequently in a database are known as frequent patterns and they must satisfy a minimum support percentage. For a term set denoted by X in document d, the covering set of X in d is represented as [X] and it includes all paragraphs which are found in required document ,dp that is dp E PS(d) and X E dp. ,Its absolute support is given by the number of occurrences of X in set of paragraphs PS(d). The fraction of the paragraphs that contain the pattern gives its relative support .If the absolute support is greater than minimum support, denoted by min_sup, then the termset X is a frequent pattern. The concept of closed sequential pattern is used here. A pattern is known as closed sequential pattern if it does not have a super pattern with the same support count.

Two important techniques used here are Pattern Deploying and Pattern Evolving. In pattern deploying, discovered patterns are interpreted as d-patterns to enhance the performance of closed patterns. Supports are evaluated for each term. Terms are weighted according to their appearance in patterns. A composition operation is used which adds the term supports within the patterns. In pattern evolution, reshuffling supports of terms within normal forms of d-patterns based on negative documents in the training set are performed. This technique is highly useful as it reduces the side effects of noisy patterns caused by the low-frequency problem. This technique is called inner pattern evolution, because within the pattern it only changes a pattern's term supports. A threshold is usually used to classify documents into relevant and irrelevant categories. Finally a shuffling algorithm is used to tune term support within discovered patterns. The time complexity of Algorithm is decided by the number of calls made for Shuffling algorithm and the number of times composition operation is used. A special variable known as offender is used and it is a pattern that carries a negative noise document. A different strategy is used in this algorithm for each type of offender. Complete conflict offenders are removed and it indicates that they can be discarded for preventing interference from these possible noises. But here a common support count is used at all levels. The situation may become worse when the database contains lots of long transactions or long high utility itemsets.

VI.UTILITY BASED PATTERN MINING

In 2013, a paper Efficient Algorithms for Mining High Utility Itemset from Transactional Databases [9] was proposed. Here profits of each term were considered.

Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. In this paper, two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets were proposed. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database

Proposed algorithms, especially UP-Growth+, not only reduce the number of candidates effectively but also out perform other algorithms substantially in terms of runtime, especially when databases contain lots of long transactions Here high utility item sets mining refers to importance or profitability of an item to users.

VII. NEURAL NETWORKS IN DATA MINING

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions [8]. The greatest breakthrough in neural network in recent years is in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications. This powerful predictive modeling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in fig.1. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non linear system that learns to perform a function from data and that adaptive phase is normally training

phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the non-linear characteristics of ANN provide it lots of flexibility to achieve input output map. Artificial Neural Networks, provide user the capabilities to select the network topology, performance parameter, learning rule and stopping criteria.

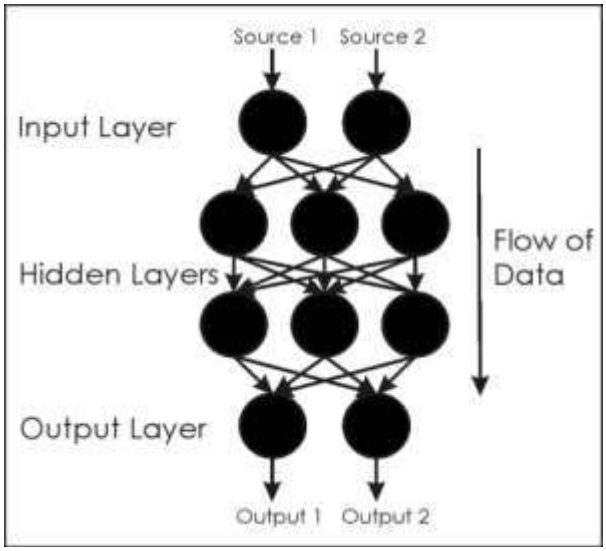

Fig: 1 Neural Network with hidden layers

VIII. GENETIC ALGORITHM IN DATA MINING

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions (schemata theory), just like nature does by combining the DNA of living beings Genetic Algorithm is basically used as a problem solving strategy in order to provide with a optimal solution. They are the best way to solve the problem for which little is known. They will work well in any search space because they form a very general algorithm. The only thing to be known is what the particular situation is where the solution performs very well, and a genetic algorithm will generate a high quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem. Genetic algorithms (GAs) are based on a biological applications; it depends on theory of evolution. When GAs are used for problem solving, the solution has two distinct stages:

1 The solutions of the problem are encoded into representations that support the necessary variation and selection operations; these representations, are called chromosomes, are as simple as bit strings.

2. A fitness function judges which solutions are the "best "life forms, that is, most appropriate for the solution of the particular problem. These individuals are favored in survival and reproduction, thus giving rise to generation.

Crossover and mutation produce a new gene individual by recombining features of their parents. Eventually a generation of individuals will be interpreted back to the original problem domain and the fit individual represents the solution.

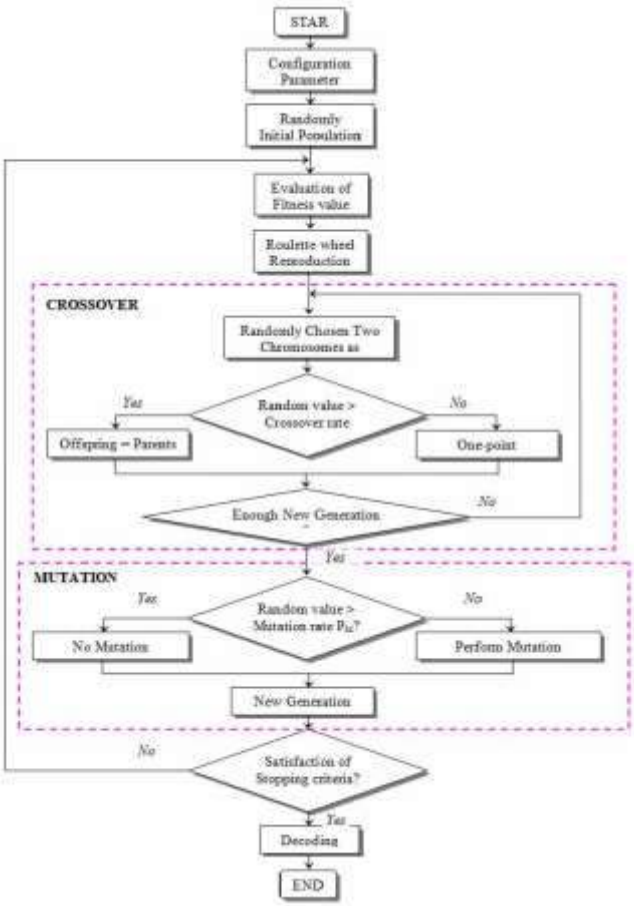

Fig 2: Structured view of genetic algorithm

IX. APPLICATION

Here some of the Data mining applications are analyzed.

1. Data Mining Applications in Healthcare

Data mining applications in health can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications.

2. Data mining is used for market basket analysis

Data mining technique is used in Market Basket Analysis. When the customer want to buy some products then this technique helps us finding the associations between different items that the customer put in their shopping buckets. Here the

discovery of such associations that promotes the business technique .In this way the retailers uses the data mining technique so that they can identify that which customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items.

3. The data mining is used as an emerging trend in the education system

In Indian culture most of the parents are uneducated .The main aim of in Indian government is the quality education not for quantity. But the day by day the education systems are changed and in the 21st century a huge number of universalities are established by the order of UGC. As the numbers of universities are established side by side, each and every day a millennium of students are enrolls across the country. With huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, and increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes.

4 Data mining is now used in many different areas in manufacturing engineering

When we retrieve the data from manufacturing system then the customer is to use these data for different purposes like finding the errors in the data ,to enhance the design methodology ,to make the good quality of the data ,how best the data can be supported for making the decision . But most of time the data can be first analyzed then after find the hidden patterns which will be controlling the manufacturing process and will further enhance the quality of the products.

5. Data Mining is used in Web usage mining

The complexity of tasks such as Web site design, Web server design, and of simply navigating through a Web site has been increasing continuously. An important input to these design tasks is the analysis of how a Web site is being used.

6. Data Mining finds its application in Text mining

Pattern mining has been used for text databases to discover trends, for text categorization, for document classification and authorship identification.

7. Data Mining is used in Bioinformatics

Pattern mining is useful in the bioinformatics domain for predicting rules for organization of certain elements in genes, for protein function prediction, for gene expression analysis, for protein fold recognition and for motif discovery in DNA sequences.

IX. CONCLUSION

There are a number of techniques evolving in the field of data mining. Some of them are Association rule mining and frequent item set mining, sequential pattern mining, and maximum pattern mining, closed pattern mining etc. But, using these techniques, discovering knowledge in the field of text mining is difficult and ineffective, because some specific useful long patterns lack required support. They possess the low-frequency problem. This paper presents an analysis of profit based mining and its advantage over other methods. Also application of Neural Networks and Genetic Algorithm in Data Mining is studied.

REFERENCES

1. *Gerard Salton and Christopher Buckley: "Term Weighting Approaches in Automatic Text Retrieval"Year-1988*
2. *R Agrawal and R Srikant : " Mining Sequential Patterns" Year-1994*
3. *X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc.Int'l Joint Conf. Artificial Intelligence (IJCAI'03), pp. 587-594, 2003*
4. *X.Yan, J.Han, and R.Afshar : Clospan: "Mining Closed Sequential Patterns in large Databases".Year-2003*
5. *Ding-Ying Chiu, Yi-Hung Wu, Arbee L. P. Chen: "An Efficient Algorithm for Mining Frequent Sequences by a New Strategy without Support Counting". Year-2004*
6. *Yuefeng Li and Ning Zhong proposed a paper named "Mining Ontology for Automatically Acquiring Web User Information Needs" Year-2006*
7. *Ning Zhong,Yuefeng Li,and Sheng-Tang Wu,"Effective Pattern Discovery for Text Mining" Ning Zhong, Yuefeng Li, and Sheng-Tang Wu Year-2012.*
8. *Ankita Agarwal,"Secret Key Encryption algorithm using genetic algorithm", vol.-2, no.-4, ISSN: 2277 128X,IJARCSSE, pp. 57-61, April 2012*
9. *Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, Fellow, IEEE "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases"Year- 2013*