OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Novel Data mining Technique to Discover Patterns from Huge Text Corpus

Mohamed Younis Mohmed Alzarrous [1], Mr.Surya Prakash Mishra [2]

[1, 2] *(M.Sc, Asst. Prof, in Computer Science in Department of Computer Science and Information Technology in SHIATS, Allahabad, India)*

Abstract: *Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Many techniques have been investigated on mining scenario from documents including the texts for required patterns respectively. There is a problem on dealing with this particular task for inventory patterns which are accurate. Research has done on this strategy; results have proven that this strategy is facing two problems, i.e., 1) synonymy and 2) polysemy (coexistence of many possible meanings for a word). So, in the text mining, we can use the techniques of pattern mining to find different text patterns, like co-occurring terms, frequent item-sets. Therefore now this present technique i.e., inventory pattern plays a crucial role in the investigation of the patterns. We first conduct a set of large-scale measurements with a collection of over different data sets into a database. We upload this database consisting of data sets for constructing the pattern taxonomy model, partial conflict tree and Chart. Based on the measurement results, we have proven that this technique works efficiently and effectively. It also provides good results for the implementation of task.*

Keywords: *Patterns, Associations, or Relationships, Sequence patterns, Classification of text.*

I. INTRODUCTION

Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data. With this, Data mining with inventory pattern came into existence and got popularized. Data mining finds these patterns and relationships using data analysis tools and techniques to build models.

There are two main kinds of models in data mining. One is predictive models, which use data with known results to develop a model that can be used to explicitly predict values. Another is descriptive models, which describe patterns in existing data. All the models are abstract representations of reality, and can be guides to understanding business and suggest actions. These data provides many benefits and plays a vital role in the entire society in terms of managerial business and analyzing the market by the particular extraction respectively. Data mining is an important part of Knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Data mining is applied effectively not only in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc.

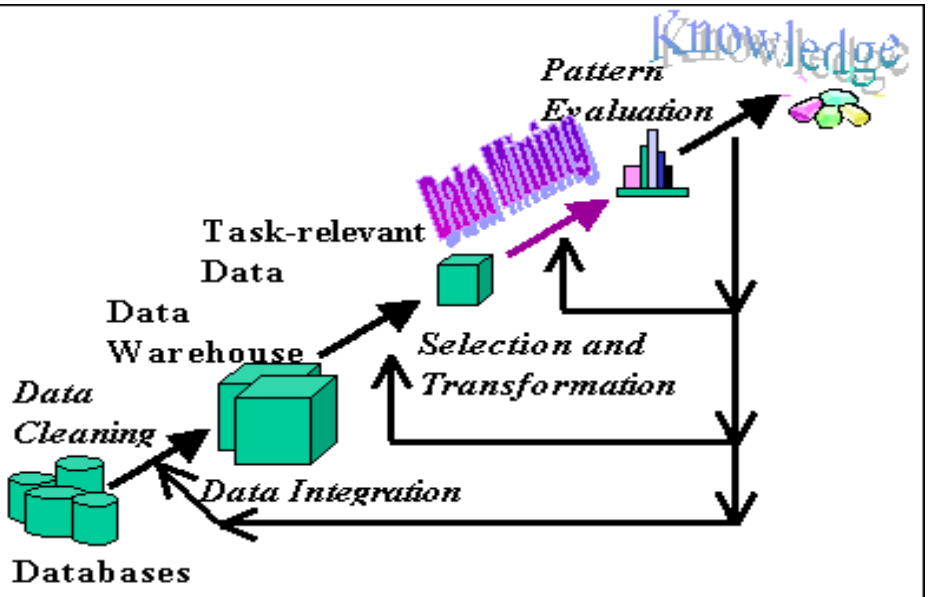


Figure 1.1: Data Mining is the core of Knowledge Discovery process

II. RELATED WORK

Necessity is the mother of invention. Since ancient times, our ancestors have been searching for useful information from data by hand. However, with the rapidly increasing volume of data in modern times, more automatic and effective mining approaches are required. Early methods such as Bayes' theorem in the 1700s and regression analysis in the 1800s were some of the first techniques used to identify patterns in data. After the 1900s, with the proliferation, ubiquity, and continuously developing power of computer technology, data collection and data storage were remarkably enlarged. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms in the 1950s, Decision trees in the 1960s and support vector machines in the 1980s.

However, a well known one is a bag of words that uses keywords (terms) as elements in the vector of the feature space. Here the Rocchio system is used to improve performance. Also many other weighting schemes were given. The problem of the bag of words approach was how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency.

Data mining techniques have been used for text analysis by extracting co-occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had lower consistency of assignment and lower document frequency for terms.

Pattern mining has been extensively studied in data mining communities for many years. A variety of algorithms such as Apriori-like algorithms, PrefixSpan, FP-tree, APADE, SLPMiner and GST have been proposed. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from large data collection. However, searching for useful and interesting patterns and rules was still an open problem.

III. Proposed System

Even though after generating the patterns there is a problem in mining, they lend themselves discovering frequent item sets and the order they appear. So to overcome all these problems a particular method has been implemented which plays a key and crucial role in the field of mining of text in the form of extraction of the data accurately followed by a good updating process in the entire phenomena. Mining of text is the detection of knowledge based on the interest in the documents consisting of text respectively. Now there is a major challenging task involved in it where it must provide service for the user in the form of accurate extraction of the information in the mining of text in the form of generated patterns. There are large numbers of experimental analysis taken place in the collection of the data and also the retrieval of the text accurately

and precisely depending on the user's choice. Therefore our present methodology overcomes all the above problems accurately and carries the success.

IV. Problems in the Existing System

For the extraction of the information and knowledge, a large number of investigations had taken place in the mining of the data previously. There are some existing techniques that follow mining by the rule of association, Sequential patterns, mining by the help of item set phenomena, Mining by clustering, Mining by the classification, mining by the Prediction. There are some limitations to these presented techniques.

- They are implemented under one particular frame of time, i.e., under the restricted areas.
- Missing data; which poses a big hurdle in this current system
- The data is very large. So, it's not always very accurate
- Quality of the data is poor

V. Solutions to These Problems

To improve the performance in the mining of a text based aspects in the form of the patterns related to the closed strategy an extra information is used in the system for this purpose we are supposed to hold that particular phenomena by the name as D pattern respectively which is mainly used for the weight evaluation. Therefore these D patterns play a major role in the form of the technique by the name term in which there is no synonym for the greater value. Therefore this particular strategy is completely differed from the generalized scenario basis where it is completely inter dependent on the terms involved in the documentation in it. Further improving the performance of the system in the form of affectivity in the mining pattern of taxonomy where the differentiation of the algorithm takes place by the name of the mining based on SP method respectively in order to observe the relatively ordered patterns where the space of searching is got negotiated. Now moving towards the reciprocal documents that is the opposite of the normal scenario based documents where the reshuffling involves in it and also which provides support by the pattern D in the original form. In order to this a patterns generated by the noise gets nullified due to the less frequency oriented data and is termed as the evolution of the inner patterns. Here the consideration takes only with respect to the inner phenomena where as the outer phenomena got escaped from the system. Now here the major concern is similar and non similar data has to be differentiated therefore in order to separate between themselves, conditionality has to be set by the name of the threshold respectively. The defining of the threshold takes place by the help of the initialization of the setting D pattern. And therefore the threshold function is represented by the following equation

$$Threshold(D\ ') = \min_{p \in D} \left(\sum_{(t,w) \in \beta(p)} support(t) \right).$$

VI. Benefits

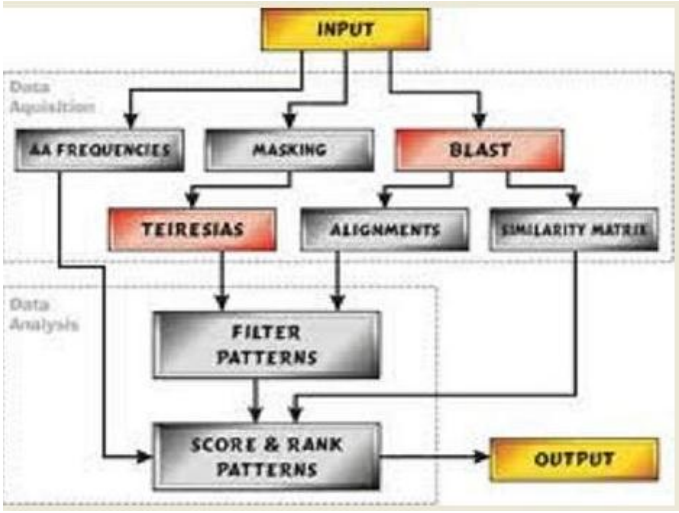
Using the D-pattern method, the text has the following benefits:-

- Improves the effectiveness by effectively
- Significantly improves the performance of information filtering
- Noise generated by the patterns are nullified due to less frequency data
- Provides accurate data
- Reduces the searching space

VII. Expected Result

In the present section by the evolution of patterns in the model of taxonomy, experimental results of PTM approach are presented and analyzed. The tabular column shows how frequently patterns occur in covering set and also displayed in the figure. As already mentioned earlier that mining of the data based on the item set where it is satisfied with rapid generation but it suffers from the implementation/ evaluation respectively. Here the comparison takes place by the experts and provides the data results accurately and in the much more efficient manner[3][4][6]. Our present method plays a crucial role on mining based on the terms and also on pattern dependent mining. and some of them includes support vector machine and the state of art.

Effective Inventory pattern



The above figure describes the step by step process of how the input has been converted to output i.e., inventory patterns.

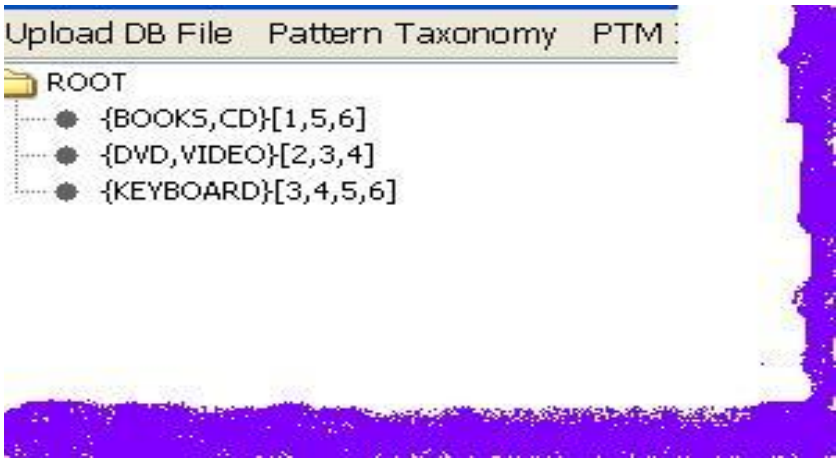
Pattern Taxonomy model:

Upload DB File Pattern Taxonomy PTM IPE

Frequent Pattern	Covering Set	Support
BOOKS,CD	P1,P5,P6	3.0
DVD,VIDEO	P2,P3,P4	3.0
BOOKS	P1,P5,P6	3.0
CD	P1,P5,P6	3.0
DVD	P2,P3,P4	3.0
VIDEO	P2,P3,P4	3.0

After uploading the database, we will get the above figure, which shows how frequent the patterns appear in covering set i.e., set of paragraphs and also shows the support value of each pattern.

Conflict tree:



The above figure shows how set of different items appear in different paragraphs

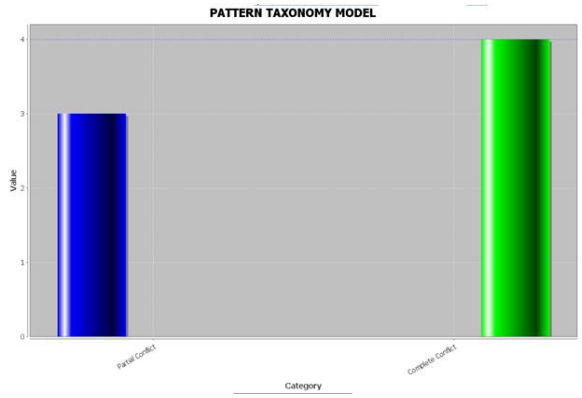


chart :

The above figure shows how the partial conflict differs from the complete conflict

VIII. Conclusion

Research has been done on this particular problem. Many techniques have been involved before this methodology and some of them are mining by association, Sequential mining of the patterns, Pattern maximization, closeness of Pattern etc. The study of the above research oriented concepts involves lot of effort and is a huge tedious job in the stream of mining text and lacks efficiency and is less effective. In addition to the above problem it has the frequency issues. It has low frequency components. Low frequency means patterns generated with most of them are small and ineffective. In order to overcome these problems a new technique is implemented for studying low frequency data and also for studying mismatched problems respectively. The technique which is proposed in this paper deals with the evolution and deployment of the patterns in the mining of text. Practical approaches confirm that it not only used for the data text but also used for the state-of-art that is the support vector machine respectively.

REFERENCES

[1] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[2] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[3] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[5] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[6] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.

[7] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.

[8] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile,"Kernel Methods for Document Filtering," TREC, trec.nist.gov/pubs/trec11/papers/kermit.ps.gz, 2002.

Name:- Mohamed Younis Mohmed Alzarrous.
Gender :- male .
date of birth :- 01/12/1983
nationality of birth :- Libya .
present nationality :- Libya .
pervious certificate :- B.Sc in computer science - Libya - tragen .
current certificate :- M.Sc in computer science -India -Allahabad .
E-mail: moh_you83@yahoo.com

Name:-Mr. Surya Prakash Mishra .
gender :- male .
date of birth :- 05/07/1982
nationality of birth :- India .
present nationality :- India .
pervious certificate :- M.C.A in computer - UP. Tech University .
current certificate :- pursuing Ph.D
E-Mail :- surya.mishra@shiats.edu.in