ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

REVIEW ON TEXT MINING WITH PATTERN DISCOVERY

Rupali Bhaisare [1], T. Raju Rao [2]

P.G. Student, Department of Computer Science & Engineering, Abha Gaikwad-Patil College of Eng., Nagpur, India[1]

Professor, Department of Computer Science & Engineering, Abha Gaikwad-Patil College of Eng., Nagpur, India[2]

Abstract: In text documents data mining techniques have been proposed for mining useful patterns. But there are some questions, how to effectively use and update discovered patterns is still an open research issue, especially in the text mining. So most existing text mining methods adopted term-based approaches but they all suffer from the problems of polysemy and synonymy. Polysemy is the word which giving the multiple meaning of word and synonymy is the word which giving the similar meaning of word. After some years, people have been adopted pattern based approaches should perform better than the term-based approaches. This paper with proposed system implements innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information with effective patterns as per the users requirements. In this paper user also getting the meaningful information without misinterpretation problem.

Keywords: Text Mining, Data Mining, Text Classification, Pattern Mining, Information Filtering, Clustering, Association Rules.

I. INTRODUCTION

In the past years, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. There is rapidly growth of digital data made available in recent years; knowledge discovery and data mining have work together a great deal of attention with an need for meaningful data into useful information and knowledge. Data mining is therefore an efficient step in the process of knowledge discovery in databases. Text mining is used to finding relevant & interesting information from huge database. Text mining is to exploit information contained in textual documents in various ways including discovery of patterns, association among entities, etc. With this paper in proposed system I 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 advantages of term based methods include efficient computational performance as well as theories for term weighting.
This paper proceeds as follows. In the next section, the background study is described. Section 3 describes related works in this field, etc.

II. BACKGROUND STUDY

A. Data Mining

Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is primarily used today by companies with a strong consumer focus on retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. And it helps to company to determine the impact on sales, customer satisfaction, and corporate profits.
Data mining software analyzes relationships and patterns in stored transaction database.

ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

B. Text Classification

Text classification is the process of classifying documents into predefined categories based on their content. This paper presents a new algorithm for text classification using data mining that requires fewer documents for training. Text classification is the task of assigning predefined categories to free-text documents. Text classification is used in several fields such as patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on.

C. Pattern Mining

Patterns are item sets, subsequence, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent item set. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a frequent pattern .A substructure can refer to different structural forms, such as sub graphs, sub trees, or sub lattices, which may be combined with item sets or subsequence.

D. Information Filtering

Information filtering system is a system that removes redundant or unwanted information from an information stream using automated or computerized methods to user. Its main goal is the management of the information overload and increment of the semantic signal-to-noise ratio. In this the user's profile is compared to some reference characteristics. Information filtering is usually works by specifying character strings, if they matched, then it indicate undesirable content that is to be screened out.

E. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups . It is a main task of exploratory or retrieving data from data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, , image analysis, pattern recognition information retrieval. Work with clustering helps to modify data pre-processing and model parameters until the result achieves the desired properties.

F. Association Rules

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.

III. RELATED WORK

1. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task [1]

In this paper the author was study the properties of phrasal and clustered indexing languages on a text categorization task, enabling us to study their properties in isolation from query interpretation issues. In this paper author improve the

ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

statistics properties of a text representation since most pairs of words tend to co-occur in one particular synchronous relationship.

2. Fast Algorithms for Mining Association Rules in large databases [2]

In this paper the author said that the problem of discovering association rules between items in a large database of sales transactions. For solving this problem author also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid algorithm.

3. Kernel methods for document filtering [3]

In this paper author said that algorithms implemented by KERMIT IST European project is concerned with the investigation of kernel methods for applications related to the categorization, retrieval, clustering and ranking of text documents and of images. Author overcomes the problem with lack of improvement in performance when polynomial kernels of degree higher than one or radial basis function kernels.

4. Mining Generalized Association Rules [4]

Author introduces the problem of mining generalized association rules with a large database of customer transactions, where each transaction consists of a set of items, and taxonomy on the items. Here solution to the problem is to replace each transaction with an "extended transaction" that contains all the items in the original transaction.

IV. METHODOLOGY

In proposed system I will make the three modules.

1. Admin.
2. Manager.
3. User.

In proposed system Admin module giving the control on whole system and also provide the help to user through manager. But user can't directly talk with admin. User can directly talk with manager for accessing the database. In proposed system only authorized (registered) user doing the searching operation with their choice of keyword. The keyword may be any alphabet, word, phrase or sentence. Then after this, keyword enters into the database and three methods apply on that keyword. When complete the process with three methods it will provide the list of data to user as per their searching keyword and save the data for future work which is shown in fig. 2.

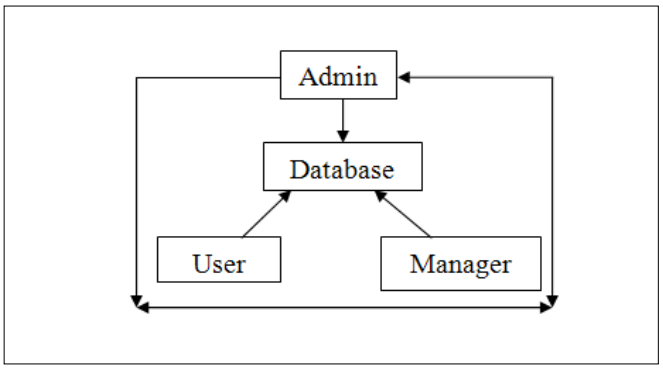

Fig.1. Proposed System Module diagram

ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

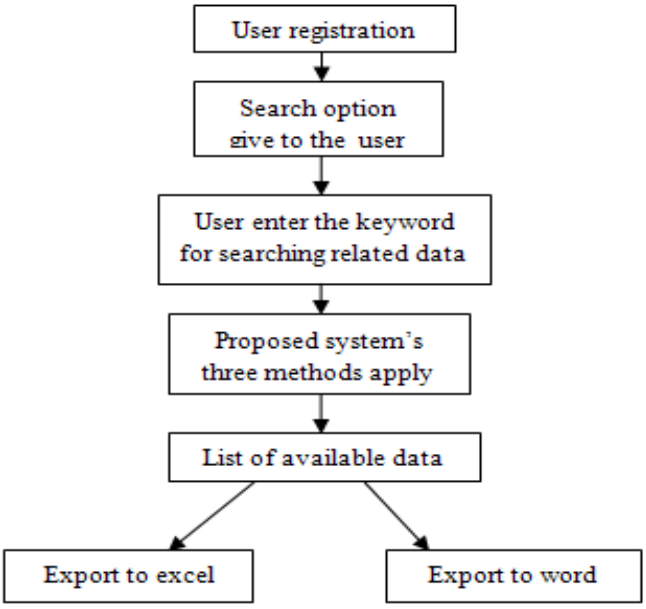

Fig.2. Data flow diagram of Proposed System

V. RESULTS

I know this is one of the most important sections which will decide whether the work is publishable or not related this paper but I am not exactly saying here because I have going to research on this topic i. e. text mining and work is not done completely. I am not giving any result exactly but in proposed system I will overcome the problem with low frequency and misinterpretation which are present in existing system.

VI. CONCLUSION

In the existing system some techniques of clustering algorithms included with some techniques such as association rule mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. But they don't give efficient meaningful information to user. And they also difficult and ineffective during used. The reason is that some useful long patterns with lack in support i.e., the low-frequency problem and misinterpretation problem. Misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. In proposed system I will use some existing algorithms included extra functions and features also. And they also giving the user friendly functionality to user for updating patterns and suggestions as per users requirements. So in proposed system some new clustering algorithms and techniques to make possible working with effective pattern discovery to overcome the low frequency and misinterpretation problems for text mining. With this conclusion the result is, the user getting the meaningful and similar information which they need from huge amount of database in faster manner with some excellent mining algorithms and techniques.

ACKNOWLEDGEMENT

I would like to thanks my guide Prof. T. Raju Sir to giving me their excellent knowledge, insightful comments and suggestions for making this paper. And guide me how I doing the preparation and search related this paper.

ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

REFERENCES

[1] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), pp. 37-50, 1992.

[2] 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.

[3] 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.

[4] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. 21th Int'l Conf. Very Large Data Bases (VLDB '95), pp. 407-419, 1995.

[5] 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.

[6] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.

[7] 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.

[8] Y. Li, W. Yang, and Y. Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 953-958, 2006.

[9] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans. Mining (ICDM '03), pp. 593-596, 2003.

[10] Y. Li and N. Zhong, "Interpretations of Association Rules by Granular Computing," Proc. IEEE Third Int'l Conf. Data Mining (ICDM '03), pp. 593-596, 2003.

[11] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.