Efficient Algorithms for Mining the Concise and Lossless Data Records of High Utility Item Sets

Mr. Sagar M. Kalbande
Computer Science & Engineering
Prof. Ram Meghe Institute of Technology & Research
Badnera.

Prof. Ms. R.R. Tuteja
Computer Science & Engineering
Prof. Ram Meghe Institute of Technology & Research
Badnera

ABSTRACT- **Data mining used to extract interesting correlations, frequent patterns, associations among sets of items in the transaction databases or other data repositories. These rules are mostly used in various areas such as telecommunication networks, market and risk management, and inventory control and so on. Wu et al. introduced the closer concept to high utility item sets. They called the extracted item sets as closed+ high utility item sets. Our approach Research project selection is an important task for government and private research funding agencies. When a large number of research proposals are received, it is common to group them according to their similarities in research disciplines. To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency is proposed. To analyze each concept at the document level, the concept based term frequency , the number of occurrences of a concept (word or phrase) c in the original document, is calculated. To extract concepts that can discriminate between documents, the concept-based document frequency , the number of documents containing concept , is calculated. After the research proposals are classified by the discipline areas, the proposals in each discipline are clustered using the text-mining technique. This step uses an algorithm to cluster the feature vectors based on similarities of research areas. The algorithm is a typical unsupervised learning neural network model that clusters input data with similarities. Finally generate summary of research document using generated cluster in each research area**

Keywords- ***Data Mining, Frequent itemset, closed+ high utility itemset, lossless and concise representation..***

I. INTRODUCTION

Text mining is a method that discovers interesting information in text documents. It is a challenge to find accurate feature in text documents to help users to find what they want. Developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets. Pattern mining techniques are used for finding appropriate features in both relevant and irrelevant documents. Pattern mining has been extensively studied in data mining communities for many years. This paper, discusses briefly five methods, i.e Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, relevant feature discovery model. The text classification is dimensionality of feature vector which is usually huge. Therefore, developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets.

II. RELATED WORK:

Many text mining methods has been implemented over the last decades. There are different methodologies that are implemented for mining text document i.e Concept-based mining model, Text-driven D-matrix method, Ontology based Text Mining Method, PPSGEN method and Varifocal Reader method. Concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure, as depicted in Fig. 1

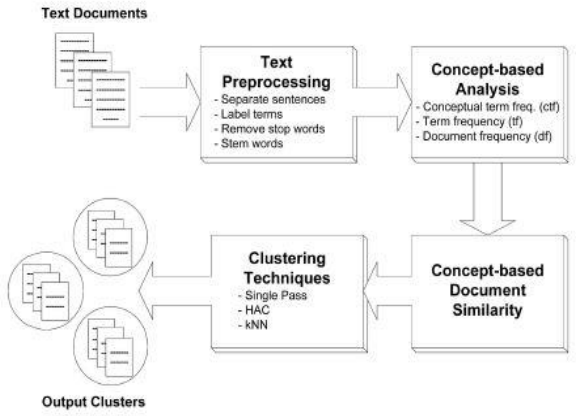


Fig. 1. Concept-based mining model system.

In FIM, to reduce the computational cost of the mining task and present fewer but more important patterns to users, many studies focused on developing concise representations, such as free sets , non-derivable sets , maximal item sets and closed item sets. These representations successfully reduce the number of item sets found, but they are developed for FIM instead of HUI mining. Frequent item set mining approach may not always satisfy a sales manager's goal. Because a retail businessman may be interested in identifying its most valuable customers (customers who contribute a major fraction of the profits to the business). The limitations of frequent item set mining motivated researchers

The study on mining of text documents discusses the most relevant mining techniques developed in recent years. Concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure [4]. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only. A new concept based mining model composed of four components, is proposed to improve the text clustering quality. The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. This method has drawback that it cannot work on web documents [5].

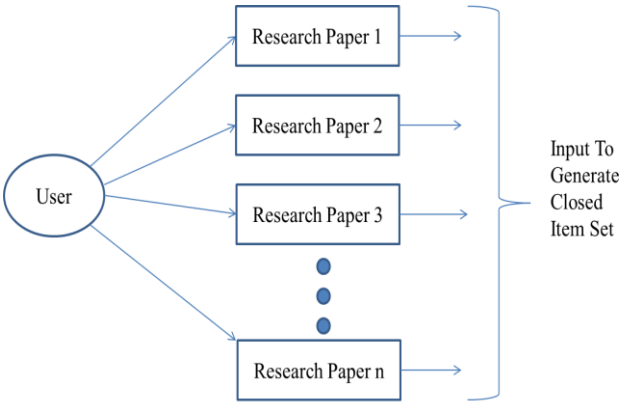
High Utility itemset Mining

Considering items quantities in transactions and their individual importance, high utility itemset mining (HUIM) received a considerable research attention [7]. Yao et al. proposed a mathematical model of utility mining by generalizing the share-confidence model. As utility mining does not fulfill the *downward closure property*, Liu et al. [11] proposed the two phase algorithm that uses the *transaction-weighted downward closure property* to prune the candidate high utility itemsets in the first phase and then all the complete sets of high utility itemsets are obtained in the second phase. To reduce the number of candidate itemsets in the first phase, Li et al. [10] also proposed an isolated items discarding strategy (IIDS) to the level-wise utility mining method. IIDS Discard isolated items and their actual utilities from transactions and transaction utilities of the database.

III. PRAPOSE WORK

A) Input The Research Base Paper Collection from User.

Research project selection is an important task for government and private research funding agencies. When a large number of research proposals are received, it is common to group them according to their similarities in research disciplines. Funding agencies such as the NSFC maintain a directory of discipline areas that form a tree structure. As a domain ontology , a research ontology is a public concept set of the research project management domain. The research topics of different disciplines can be clearly expressed by a research ontology.

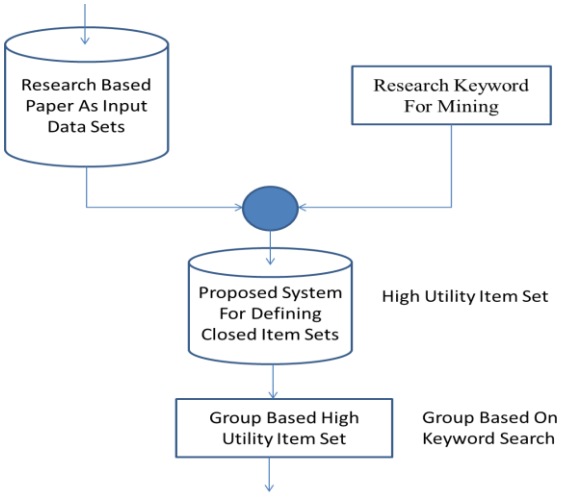


Paper As Input For Mining

B) Processing The Research Base Paper With Keyword Found.

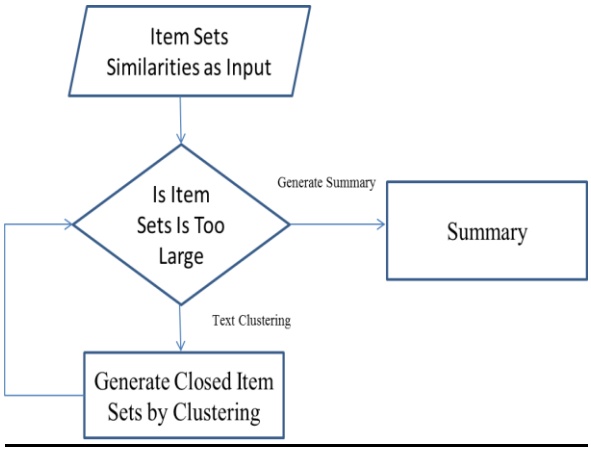
To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency is proposed. Sentence importance assessment is one of the two key steps , which aims to assign an importance score to each sentence in the

given paper. The score of each sentence will be used in the summarization process. In this study, we introduce a few useful features and High utility item sets .Generate group based on keyword search.



C) Summaries The research keyword data

After the research proposals are classified by the discipline areas, the proposals in each discipline are divide into item sets using the text-mining technique. This step uses an Apriori algorithm to sets the feature vectors based on similarities of research areas. The Apriori algorithm is a typical unsupervised learning neural network model that item sets input data with similarities. Finally generate summary of research document using generated cluster in each research area.



IV. CONCLUSION

The text classification is dimensionality of feature vector which is usually huge. Therefore, developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets. Typically feature extraction method aims to convert the original high-dimensional data set to a lower-dimensional data set by projecting process using algebraic transformations. In this paper, we addressed the problem of redundancy in high utility itemset mining by proposing a lossless and compact representation named closed+ high utility itemsets

Acknowledgment

I wish to thank my guide prof. R.R.Tuteja madam for their valuable support, guidance and suggestions.

REFERENCE

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.

[2] S. Kannimuthu , Dr. K. Premalatha iFUM - Improved Fast Utility Mining International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011.

[3] J.-F. Boulicaut, A. Bykowski, and C. Rigotti,"Free-sets: A condensed representation of Boolean data for the approximation of frequency queries," Data Mining Knowl. Discovery, vol. 7, no. 1, pp. 5–22, 2003.

[4] T. Calders and B. Goethals, "Mining all non-derivable frequent itemsets," in Proc. Int. Conf. Eur. Conf. Principles Data Mining Knowl. Discovery, 2002, pp. 74–85.

[5] S. Kannimuthu , Dr. K. Premalatha iFUM - Improved Fast Utility Mining International Journal of Computer Applications (0975 – 8887) Volume 27– No.11, August 2011.

[6] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient mining of association rules using closed itemset lattice," J. Inf. Syst., vol 24, no. 1, pp. 25–46, 1999.

[7] C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong, and Y.-K. Lee. Efficient tree structures for high utility pattern mining in incremental databases. *Knowledge and Data Engineering, IEEE Transactions on*, 21(12):1708–1721, 2009.

[8] C. Lucchese, S. Orlando, and R. Perego, "Fast and memory efficient mining of frequent closed itemsets," IEEE Trans. Knowl. Data Eng., vol. 18, no. 1, pp. 21–36, Jan. 2006.

[9] H.-F. Li, H.-Y. Huang, Y.-C. Chen, Y.-J. Liu, and S.-Y. Lee. Fast and memory efficient mining of high utility itemsets in data streams. In *Data Mining, 2008. ICDM '08. Eighth*

IEEE International Conference on, pages 881–886, 2008.

[10] Y.-C. Li, J.-S. Yeh, and C.-C. Chang. Isolated items discarding strategy for discovering high utility itemsets. *Data & Knowledge Engineering*, 64(1):198 – 217, 2008.

[11] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499..

[12] K. Gouda and M. J. Zaki, "Efficiently mining maximal frequent itemsets," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 163–170.

[13] Chao Gao, Xin Zhang, and Hui Wang, "A Combined Method for Multi-class Image Semantic Segmentation", *IEEE Transactions on Consumer Electronics*, Vol. 58, No. 2, PP. 596-604, May 2012.

[14] Torben Pätz and Tobias Preusser, "Segmentation of Stochastic Images With a Stochastic Random Walker Method," *IEEE Transactions On Image Processing*, VOL. 21, NO. 5, PP. 2424-2433, MAY 2012.