Impressive pattern discovery to provide Effective Output through Text Mining

Dr.M.V.SIVA PRASAD Phd , P.SANDEEP REDDY M.Tech, S.TIRUMALARAO B.Tech

#ANURAG ENGINEERING COLLEGE,KODAD,ANDHRA PRADESH

Abstract

Due to the rapid growth in digitalization, there exist huge amount of digital text data. In order to get useful information from this data there is a great need for mining techniques. Text mining is nothing but getting the useful information of the given text. Most of the existing text mining techniques are based on term frequency. The problem using this term based techniques is, that they will consider only the frequency of the terms in the given text, but useful terms may have less frequency and less useful terms may have high frequency, so we can't get the useful terms if our support count is high. In other way if the support count is less, there is a chance to get much amount of useless data. To overcome this problem, here we are going to use pattern based approach.

The Pattern based mining approach is pretty much better than the term based mining approach in performance. The proposed pattern based mining approach is a innovative and a efficient pattern discovery technique that 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. For evaluating the proposed method we used Reuter's data set.

I. INTRODUCTION

Text mining is that the discovery of fascinating knowledge in text documents. it's difficult issue to seek out correct data in text documents to help users to seek out what they require. Many applications, like marketing research and business management, will profit by the utilization of the information and data extracted from an oversized amount of knowledge. data discovery will be effectively use and update discovered patterns and apply it to field of text mining . Data mining is so an important step within the method of knowledge discovery in databases, which suggests data mining has all strategies of data discovery method and presenting modeling section
that is application of strategies and rule for
calculation of search pattern or models. within the past decade, a major variety of knowledge mining techniques are bestowed so as to perform different data tasks. These techniques embrace association rule mining, frequent item set mining, sequential pattern mining, most pattern mining and closed pattern mining. Most of them area unit proposed for the aim of developing economical mining algorithms to search out explicit patterns at intervals a reasonable and acceptable time-frame . With an oversized variety of patterns generated by victimization the data mining approaches, a way to effectively exploit these patterns remains Associate in Nursing open analysis issue. Text mining is that the technique that helps users notice helpful data from an oversized quantity of digital text information . it's thus crucial that a decent text mining model ought to retrieve the information that users need with relevant potency. ancient data Retrieval (IR) has an equivalent objective of mechanically retrieving as several relevant documents as doable while filtering out digressive documents at an equivalent time. However, IR-based systems don't adequately provide users with what they actually need. several text mining ways are developed so as to achieve the goal of retrieving for data for users. we tend to concentrate on the event of a information discovery model to effectively use and update the discovered patterns and apply it to the sector 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 additionally referred to as as data discovery in databases as a result of, we often notice in literature text mining as a method with series of partial steps among alternative things additionally data extraction also because the use of information mining. after we analyze knowledge in data discovery in databases is aims of finding hidden patterns also as connections in those knowledge. whereas the flexibility to search for keywords or phrases in an exceedingly assortment is currently widespread such search solely marginally supports discovery as a result of the user should take the words to appear for. On the opposite hand, text mining results will recommend "interesting" patterns to appear at, and therefore the user will then settle for or reject these patterns as fascinating. during this analysis we tend to gift pattern taxonomy model that extracting descriptive frequent patterns by pruning the nonsense ones. patterns ar sorted supported their repeations.

II. IMPLEMENTATION

The main objective of this work is to find the patterns in the given input files. After finding the patterns, classification will be done using this patterns.Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.

In order to solve the above paradox, this project presents an effective pattern discovery technique, which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and try to reduce their influence for the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evolution. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents. The following figure illustrate steps involved in finding the patterns

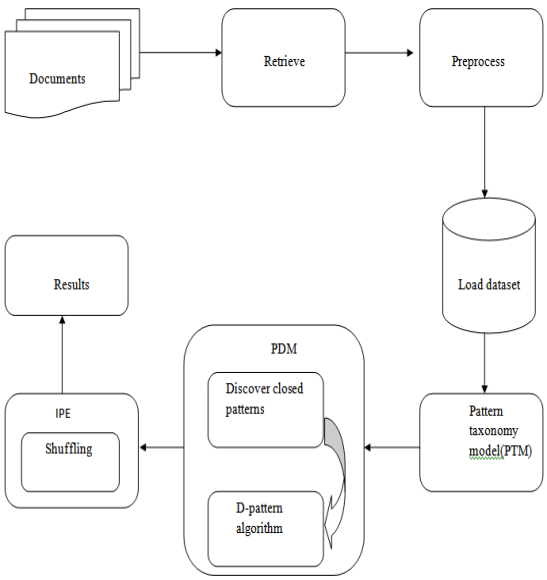

Fig 3.1: System Architecture

First of all we will retrieve some documents from document set and then we will perform
preprocessing, then we will store this preprocessed data in data base or temp files. Then we will apply Pattern Taxonomy Model for this. After PTM we will do Pattern Deploying Method (PDM), includes Discover Closed Patterns and D_Pattern algorithm. After PDM we will do Shuffling, this will give final result that means patterns.

III. Module Description

Based on the system architecture we can divide the proposed system into the following modules

- Frequent and Closed Patterns
- Closed Sequential Patterns
- D-Pattern Mining Algorithm
- Inner Pattern Evolution
- Classification

Description of above mentioned modules is given as,

3.1.1 Frequent and Closed Patterns

In this module given a term set X in document d, X' is employed to denote the covering set of X for d, which incorporates all paragraphs displaced person happiness to Paragraph Set. Its absolute support is that the range of occurrences of X in notation. Its relative support is that the fraction of the paragraphs that contain the pattern, that is, supr. A term set X is termed frequent pattern if its supr or supa is bigger than or adequate a minimum support. The duplicate terms were removed. All the Frequent patterns might not be helpful, hence, we tend to believe that the shorter one may be a noise pattern and expect to stay the larger pattern solely. Given a term set X, its covering set X' may be a set of paragraphs. Similarly, given a collection of paragraphs notation we will outline its term set. The closure of X is outlined. A pattern X additionally a term set is termed closed if and as long as X is closed. Patterns is structured into a taxonomy by victimization the is-a (or subset) relation, wherever the nodes represent frequent patterns and their covering sets; non closed patterns is pruned; the perimeters area unit “is-a” relation. when pruning, some direct “is-a” retaliations is also modified. Smaller patterns within the taxonomy area unit sometimes a lot of general as a result of they might be used oft in each positive and negative documents; and bigger patterns. The linguistics info are employed in the pattern taxonomy to boost the performance of victimization closed patterns in text mining.

3.1.2 Closed Sequential Patterns

In this module a ordered pattern <t1; . . . ; tr> is an ordered list of terms. A sequence s1 <x1; . . . ; xi> could be a subsequence of another sequence s2<y1;

. . . ; yj>. . Given s1 v s2, we tend to sometimes say s1 could be a sub pattern of s2, and s2 could be a super pattern of s1. Given a pattern an ordered term set X in document d, X' continues to be wont to denote the covering set of X, which incorporates all paragraphs note. Its absolute support is that the variety of occurrences of X in note that's supa. Its relative support is that the fraction of the paragraphs that contain the pattern, that is, supr. A ordered pattern X is termed frequent pattern if its relative support or absolute support is bigger than or up to a minimum support. The property of closed patterns will be wont to outline closed ordered patterns.

3.1.3 D-Pattern Mining Algorithm

In this module to enhance the potency of the pattern taxonomy mining, associate formula, SPMining, is employed to seek out all closed successive patterns, that used the well-known Apriori property so as to cut back the looking out area. formula is employed to explain the coaching method of finding the set of d-patterns. for each positive document, the SPMining formula is 1st known as giving rise to a collection of closed successive patterns SP. the most focus of this project is that the deploying method, that consists of the d-pattern discovery and term support analysis. In formula all discovered patterns during a positive document area unit composed into a dpattern giving rise to a collection of d-patterns stateless person. Thereafter, term supports area unit calculated supported the traditional forms for all terms in dpatterns. Let m be the amount of terms in T, n be the amount of positive documents during a coaching set, K be the common range of discovered patterns during a positive document, and k be the common range of terms during a discovered pattern.

3.1.4 Inner Pattern Evolution

In this module reshuffle is employed to support of terms among traditional sorts of d-patterns supported negative documents within the coaching set. The technique are helpful to scale back the facet effects of clamant patterns attributable to the low-frequency drawback. this method is named inner pattern evolution here, as a result of it solely changes a pattern's term supports among the pattern. A threshold is typically accustomed classify documents into relevant or unsuitable classes. so as to scale back the noise, we want to trace that d-patterns are accustomed produce to such a blunder. we tend to decision these patterns offenders of nd. Associate in Nursing bad person of nd may be a d-pattern that has a minimum of one term in nd. There ar 2 varieties of offenders, a whole conflict bad person that may be a set of nd; and a partial conflict bad person that contains a part of terms of nd. the essential plan of change patterns is explained as follows: complete conflict offenders ar faraway from d-patterns initial. For partial conflict offenders, their term supports ar reshuffled so as to scale back the consequences of noise documents. the most method of inner pattern evolution is enforced by the algorithmic program IPEvolving. The inputs of this algorithmic program ar a group of d-patterns refugee, a coaching set D. The output may be a composed of d-pattern. The algorithmic program is employed to estimate the edge for locating the noise negative documents. It revise term supports by victimization all noise negative documents. It additionally notice noise documents and also the corresponding offenders. Shuffling is employed to update NDP in keeping with noise documents. The task of algorithmic program Shuffling is to tune the support distribution of terms among a d-pattern. a unique strategy is devoted during this algorithmic program for every style of bad person. within the algorithmic program Shuffling, complete conflict offenders ar removed since all parts among the d-patterns ar control by the negative documents indicating that they'll be discarded for preventing interference from these doable "noises."

IV. RESULTS

In this work, I took Reuters data set as input data. And also we can take simple text as input data. Initially we will do reprocessing on input data, it will remove stop words. Then we will find frequent patterns and in order to exclude negative terms, we will apply pattern taxonomy techniques, IP Evolving techniques. Because of these techniques finally we will get only positive patterns

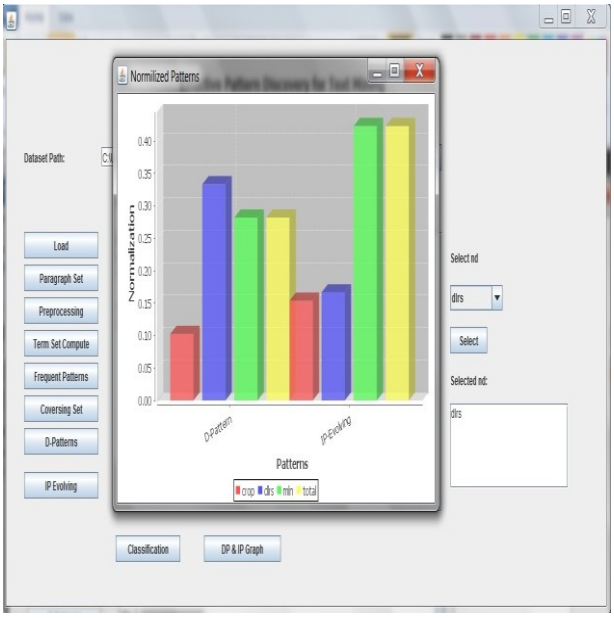

Fig 5.1 Sample Screenshot showing the weights before and after IP Evolution

The above figure shows the d-pattern and Inner Pattern Evolution graph. That means it shows the weights of D-patterns before Inner Pattern Evolution and after Inner Pattern Evolution.

CONCLUSION

Many data processing techniques are planned within the last decade. These techniques embody association rule mining, frequent itemset mining, consecutive pattern mining, most pattern mining, and closed pattern mining. However, victimization these discovered data (or patterns) within the field of text mining is troublesome and ineffective. the rationale is that some helpful long patterns with high specificity lack in support (i.e., the low-frequency problem). we tend to argue that not all frequent short patterns area unit helpful. Hence, misinterpretations of patterns derived from data processing techniques result in the ineffective performance. during this analysis work, a good pattern discovery technique has been planned to beat the low-frequency and interpretation issues for text mining. The planned technique uses 2 processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

REFERENCES

[1] M. Sassano, "Virtual Examples for Text Classification with Support Vector Machines," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '03), pp. 208-215, 2003.
[2] X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2003.
[3] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.
[4] 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.
[5] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059- 1082, 2003
[6] 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.
[7] R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.
[8] 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
[9] 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
[10] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.
[11] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.
[12] 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.
[13] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.
[14] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637,2007

AUTHOR PROFILE

Dr. M.V. Siva Prasad was received B.E from Gulbarga University, M.Tech from VTU, Belgaum & awarded Ph.D from Nagarjuna Univeristy, Guntur. Presently Working as a Principal in Anurag Engineering College, Ananthagiri (V), Kodad (M), Nalgonda (Dt.), Andhra Pradesh, India.

P. Sandeep Reddy received Master of Technology (Computer Science & Engineering) from JawaharlalNehru Technological University (JNTUH). My research interests include Information Security, Web Services, Cloud Computing, Data Mining and Mobile Computing. Presently working as Associate Professor in CSE Department in Anurag Engineering College (AEC), Ananthagiri (V), Kodad (M), Nalgonda (Dt.), Andhra Pradesh, India.

S.Tirumala Rao Pursuing Master of Technology (Computer Science & Engineering) from Jawaharlal Nehru Technological University (JNTUH). My research interests include Information Security, Web Services, Cloud Computing, Data Mining and Mobile Computing. Presently Pursuing Master of Technology in the department of CSE in Anurag Engineering College (AEC), Ananthagiri (V), Kodad (M), Nalgonda (Dt.), Andhra Pradesh, India.

