Pattern Based Mining For Relevant Document Extraction

Priyanka R. Magar
Department of CSE
M. S. Bidve Engineering College
Latur, Maharashtra, India
priyankarmagar@gmail.com

Prof. C. S. Biradar
Department of CSE
M. S. Bidve Engineering College
Latur, Maharashtra, India
chetanbiradar22@gmail.com

Abstract— This paper presents efficient mining algorithm for discovering patterns from large text collection and search for useful and interesting patterns**.** For extracting useful information we used pattern based model containing frequent sequential patterns and pruned the meaningless patterns. Here an innovative and effective technique is used for pattern discovery which includes SPM & FP growth algorithms for pattern mining and applies the processes of pattern deploying, pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

Keywords-*Sequential pattern mining, FP Growth, Text mining, Pattern deploying, Pattern evolving.*

*****

I. INTRODUCTION

Approximately 90% world's data is held in unstructured formats. In order to get useful information from this data there is great need for mining techniques. As the rapid growth of Text data and the increasing need of a more sensible and rational search system, Text mining has gained an important status in the data mining field. Text mining, when looked upon in data mining terms, is to extract, associate and analyze the information from the Text data sources [1].

Text Mining helps in discovery of interesting knowledge from text documents. Challenging issue is to find accurate knowledge from text documents that will help users to find what they actually want. There is variation in Text Mining and Web Search. In Web search, user is looking for the information that is already known and which has been written by someone else. So the problem that arises is to push aside all the material that is currently not relevant to our needs and to find the relevant information. The goal of text mining is to discover unknown information which is not known by others and thus was not yet written down.

There are various tasks in text mining. These are text categorization, clustering, information extraction, text retrieval etc. It not only helps to find important text but also find frequency of the text. Text mining also tells the relationship between the text. There is difference between information extraction and text retrieval i.e. information extraction is to extract the piece of information from unstructured document whereas text retrieval is done before the information extraction i.e. it is the branch of information extraction and in this the important information is in the form of text. All the text which is achieved from the result will be considered and using this process the person will extract the desired information. Text retrieval is collecting all text from textual document and then analyzes the result for the details.

II. RELATED WORK

It is obvious that a Text mining system would be valuable once it is capable of quickly and accurately responding to the users needs. It's difficult issue to find out proper data in text documents to help users to seek out what they want. It is a demanding work to use those patterns and also bring them up to date. Earlier term based methods are provided by Information Retrieval (IR) techniques. The term based methods are organized into probability models [2], rough set models [3] and SVM based models [4]. All term based methods suffer from troubles such as polysemy and synonymy. When a word has a variety of meanings, it is known as polysemy. When a variety of words have the equivalent meaning, it is called synonymy. Thus the semantic meaning of various discovered terms are unpredictable for answering what users want.

Because of this reason, many years people believed that phrase based techniques are better than that of term based technique. However, the experiments in the field of data mining [5] have not been proved. The possible reason include the phrases have less properties pertaining to statistics when compared with terms; frequency of occurrence is low; noisy and redundant phrases are more. Though there are some disadvantages, the sequential patterns turn out to be capable alternatives to phrases.

In [6], statistical method called Latent Semantic Indexing (LSI) was proposed, in which implicit higher-order structure in the association of words and objects was considered that improved retrieval performance by up to 30%. In [7], describes selection functions for reducing the number of features. Various dimensionality reduction approaches based on feature selection techniques are Information Gain, Mutual Information, Chi-Square, Odds ratio. Categorization performance was good with use of proportional assignment strategy and statistical classifier.

Some researchers have used phrases instead of individual words. In [8], the combination of unigram and bigrams was chosen for document indexing in text categorization and evaluation was carried out based on variety of feature

evaluation functions (FEF). A phrase-based text representation for Web document management was also proposed in [9].

For the challenging issue, closed sequential patterns have been used for text mining in [10], which proposed that the concept of closed patterns in text mining was useful and had the potential for improving the performance of text mining. In [11], Pattern taxonomy model is used to improve the effectiveness by effectively using closed patterns in text mining. In [12], two-stage model that used both term-based methods and pattern based methods was introduced to significantly improve the performance of information filtering.

III. INFORMATION RETRIEVAL USING PATTERN BASED METHOD

The main objective of this work is to find the specific terms in the given input files using pattern based method. This method presents an effective solution for knowledge discovery technique, which can solve the problems like misinterpretation and low frequency of occurrence.

In this paper, we assume that all documents are split into paragraphs. So a given document d yields a set of paragraphs PS(d).

A. Basic definition

1) Absolute and Relative support

Given a document $d = \{S_1, S_2,\ldots, S_n\}$, where S_i is a sequence representing a paragraph in d. Let P be a sequence. We call P a sequential pattern of d if there is a $S_i \in d$ such that $P \sqsubseteq S_i$. The **absolute support** of P is the number of occurrences of P in d, denoted as $supp_a(P) = |\{ S \mid S \in d , P \in S \}|$, The **relative support** of P is the fraction of paragraphs that contain P in document d, denoted as $supp_r(P) = supp_a(P) / | d |$.

2) Frequent Sequential Pattern

The basic definition of sequences used in S.-T. Wu et al. [10] study is described as follows.

Let $T = \{t_1, t_2,\ldots, t_k\}$ be a set of all terms, which can be viewed as keywords in text datasets. A sequence $S = <s_1, s_2,\ldots, s_n>, (s_i \in T)$ is an ordered list of terms. A sequence $\alpha = <a_1, a_2,\ldots, a_n>$ is sub-sequence of another sequence $\beta = <b_1, b_2,\ldots, b_m>$ denoted by $\alpha \sqsubseteq \beta$, if there exist integers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$, such that $a_1 = b_{i1}, a_2 = b_{i2},\ldots, a_n = b_{in}$. The sequence α is a *proper* sub-sequence of β if $\alpha \sqsubseteq \beta$ but $\alpha \neq \beta$, denoted by $\alpha \sqsubset \beta$. For instance, sequence <A, C> is a sub-sequence of sequences <A, B, C>. However, <B,A> is not a sub-sequence of <A, B, C> since the order of terms is considered. In addition, we also can say sequence <A, B, C> is a super-sequence of <A, C>. The problem of mining sequential patterns is to find the complete set of sub-sequences from a set of sequences whose support is greater than a user predefined threshold, *min_sup*.

B. System Architecture

The proposed model includes two phases: the training phase and the testing phase. In the training phase, the proposed model first calls PTM to find d-patterns in positive documents (D^+) based on a min sup, and evaluates term supports by deploying d-patterns to terms. It also calls IPEvolving to revise term supports using noise negative documents. In the testing phase, it evaluates weights for all incoming documents. The incoming documents then can be sorted based on these weights. And finally gets the most relevant document.

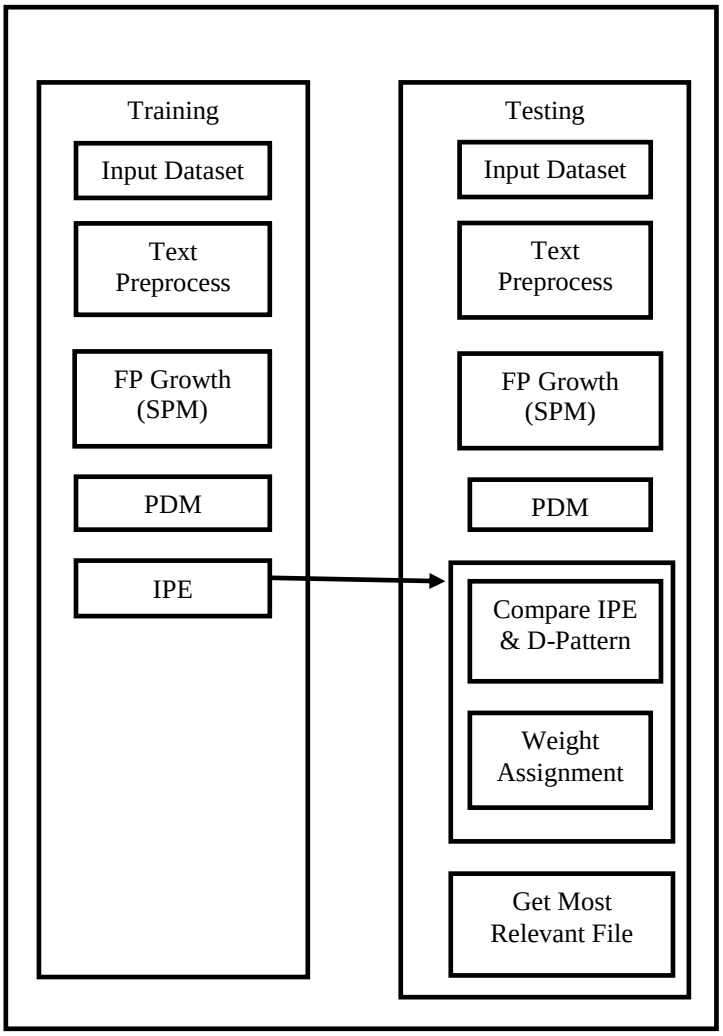

Figure 1. Block Diagram of Proposed Pattern Discovery Technique.

1) Text Preprocess

All words passes to preprocessing level. Irrelevant terms are eliminated there. This process is also called as *tokenization* process. It consists of two kinds operations such as stop list removal, stem word removal.

a) Stop List Removal: Stop words are words which are filtered out prior to, or after, processing of natural language data. They typically comprise prepositions, articles, and so on.

b) Stem Word Removal: Stemming is the process for reducing inflected words to their stem base or root form. It generally a written word forms. In this preprocess, the text documents have to be processed using the Porter stemmer. It removes the Suffix's of the words. These words are useful in the text mining for clustering the text documents. In the text mining process, we collect the documents and each documents are composed into the set of terms or words. The words have a same meaning in stem process. The suffixes of the words, singular and plural words are considered into a one single word for meaningful text clustering process.

ISSN: 2321-8169
170 – 173

2) Pattern Taxonomy Model

In PTM, documents split into set of paragraphs and each paragraph consists of set of words. At this stage, apply the data mining technique to find frequent patterns and generate pattern taxonomies. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. Also, the weight of the term which is occurring in extracted pattern is calculated. The PTM discovered frequent sequential patterns and pruned meaningless patterns in the text documents.

a) FP Growth

Mining frequent patterns in transaction databases, time-series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist long patterns.

In [13], we propose a novel frequent pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree-based mining method, FP-Growth, for mining the complete set of frequent patterns by pattern fragment growth.

Efficiency of mining is achieved with three techniques:

(1)A large database is compressed into a highly condensed, much smaller data structure, which avoids costly, repeated database scans,

(2)Our FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and

(3)A partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining confined patterns in conditional databases, which dramatically reduces the search space. Our performance study shows that the FP-Growth method is efficient and scalable for mining both long and short frequent patterns.

3) D-Pattern Mining

The main focus of the paper is deploying process which consists of- d pattern discovery and term support evaluation. The worth of patterns can be projected by assigning an evaluated value based on one existing function. In [11] pattern deploying methods are anticipated for the use of knowledge discovery. All discovered patterns are not interesting because a number of noise patterns are also extracted from the training dataset. Information from the negative sample is not demoralized during that concept learning. The negative document also contains useful information to identify ambiguous pattern in the concept. It is easier to locate the associated document if the same patterns appear in the positive document. But if the analogous pattern appears in the negative document it will be complicated. To enlarge the effectiveness it is indispensable for a system to exploit ambiguous pattern from the negative examples in order to decrease their influence.

4) IPE and Shuffling

Patten evolution is used to identify the noisy patterns in documents and update d-pattern by shuffling. This technique helps to reduce the effects of noisy patterns because of the low frequency problem. This method is called inner pattern evolution because it only changes a pattern's term supports within pattern. A threshold is used to categorize documents into relevant or irrelevant categories. In order to diminish the noise, d-patterns are tracked and find out which pattern give rise to such an error [14]. These patterns are offenders. There are two types of offenders complete conflict offenders and partial conflict offenders, the idea of updating patterns are explained here: Firstly, complete conflict offenders are deleted from discovered d-patterns then for the partial conflict offenders reshuffling of their term support.

IV. RESULTS

In this paper, we have checked the performance of the proposed system by using two parameters- time and minimum support. TABLE I. shows performance of SPM and FP Growth with respect to the execution time in ms as well as memory in MB required for two methods PTM(SPM) and FP Growth using three input text files with different sizes. Here the minimum support is same for three algorithms.

We can see that FP Growth requires less time compared to other algorithm. Hence the proposed FP Growth algorithm is time efficient.

TABLE I. PERFORMANCE OF SPM AND FP GROWTH.

Input Dataset	Algorithms Used	Execution time (ms)	Memory in MB
Education	SPM	270	7.3
	FP Growth	230	4.2
Sports	SPM	260	9.5
	FP Growth	200	4.5
Medical	SPM	270	4.5
	FP Growth	210	3

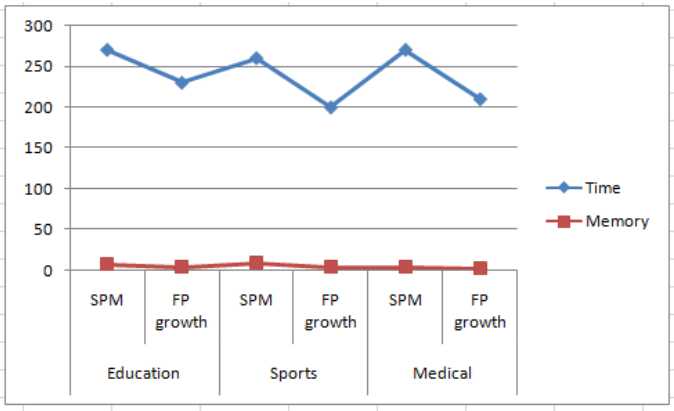

Figure2. Performance of SPM and FP Growth algorithm with different input dataset.

Figure2 shows that execution time depends on the available resources of the PC on which our tool was running. The result shows that the implemented system using FP Growth is superior than SPM (Sequential pattern mining).

V. CONCLUSION

In this paper we showed that general data mining methods are applicable to text analysis tasks. Moreover, we presented a general framework for text mining. The framework follows the general Knowledge discovery process, thus containing steps from preprocessing to the utilization of the results. Numbers of data mining techniques have been proposed in the last decade. A comprehensive comparison of data mining methods applied for text mining task is performed in this study.

In proposed technique we can take different texts file as input and applies SPM algorithm to find frequent sequential patterns. Patterns are more specific and carry more information than terms. So pattern mining based technique is used in proposed system. This system solves low frequency and misinterpretation problem. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

Our contribution is we have developed FP Growth algorithm of pattern mining. The results have shown that the execution time required by FP Growth algorithm is less than time required by SPM algorithm. Hence FP Growth is more time efficient compare to SPM. Also, we have created a database containing set of input text files to test the efficiency of our technique.

REFERENCES

[1] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, " Effective Pattern Discovery for Text Mining",IEEE Transactions on Knowledge and Data.

[2] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[3] Y. Li, C. Zhang and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[4] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.

[5] S.Scott and S. Matwin,"Feature Engineering for Text Classification," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 379-388, 1999.

[6] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers,vol. 23, no. 2, pp. 229-236, 1991.

[7] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp.212-217, 1992.

[8] 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.

[9] R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. InformationTechnology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[10] 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.

[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.

[12] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering," Proc. ACM 17th Conf. Information and Knowledge Management (CIKM '08), pp. 1023-1032,2008.

[13] 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.

[14] S.-T.Wu, Y. Li, and Y. Xu. "An effective deploying algorithm for using pattern-taxonomy" In iiWAS'05, pages 1013–1022, 2005.