International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 8 Issue 1 –APRIL 2014.

PATTERN DISCOVERY MODEL FOR EFFECTIVE TEXT MINING

R.Jayamari
ME Computer Science and Engineering, Infant Jesus College of Engineering, Tuticorin, India

***ABSTRACT*— Text mining is the extraction of interesting knowledge from text documents. A vast majority of text mining methods are based on term-based approaches which extract terms from a training set for describing relevant information. However, the quality of the extracted terms in text documents may be not high because lot of noise in text. In a various phrases that have more semantics than single words to improve the relevance, but many experiments do not support the effective use of phrases since they have low frequency of occurrence, and include many redundant and noise phrases. This paper presents an effective pattern discovery approach for text mining. This approach discovers pattern deploying and pattern evolving in text documents for identifying the most informative contents of the documents and then utilizes the identified contents to extract useful features for text mining.**

Index Terms—Text mining, text classification, pattern mining, pattern evolving, information filtering.

I. INTRODUCTION

Data mining is the Extraction of interesting information or knowledge of the data from large database. Data mining also known as the knowledge Discovery in Database. The large number of patterns generate in the data mining methods in text documents. Here specified how to effectively used in the patterns mining in text documents, it is important one how to differ from the other documents in text mining. An effective pattern discovery technique, first calculates discovered specificities of pattern and then evaluates term weights according to the distribution of documents. In weights that is how many terms occur in the particular paragraph. Calculates the terms before assume the documents are various paragraphs. Discovery pattern techniques presents on innovative and effective pattern discovery techniques in text documents. The two techniques used in this processes pattern deploying and pattern evolving. These techniques refine the discovered patterns in text documents. These patterns in text documents effectively use and modify discovered patterns an open research issue. And also reduce the problems of low-frequency and misinterpretation problems in Text mining.To overcome the disadvantages of phrase-based approaches, pattern mining based approaches proposed the concept of the closed sequential pattern and non closed sequential pattern. Pattern taxonomy model specified the closed sequential pattern and pruned the non closed patterns. These pattern mining-based approaches have shown improvements on the effectiveness of documents and extent improvements on the effectiveness. The most people think pattern-based approaches could be a significant alternative, but consequently less significant improvements are made for the effectiveness compared with term based methods.

This text mining method extract the related information from the large amount of data. 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 in documents for solving the misinterpretation problem. In this paper analyze the large amount of data used in a pattern deploying and pattern evolution methods. First, create the frequent and closed pattern in the pattern taxonomy model. Pattern Deploying Method algorithm discovered pattern in a positive documents are composed into a d-patterns. The Inner pattern evolution technique, to reduce the side effects of

noisy pattern. The noise pattern because of the low-frequency problem.

II. OVERVIEW

The 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. The one of the disadvantages of the existing system Phrases have inferior statistical properties to terms. They have low frequency of occurrence, and there are large numbers of redundant and noisy phrases among them. In our proposed system, 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.

An effective pattern discovery technique, is discovered and Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns. These patterns Solves Misinterpretation Problem . Considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries 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 incoming documents then can be sorted based on these weights. The proposed approach is used to improve the accuracy of evaluating term weights. Because, the discovered patterns are more specific than whole documents.To avoiding the issues of phrase-based approach to using the pattern-based approach.Pattern mining techniques can be used to find various text patterns.

Load the list of all documents. The user to retrieve one of the documents these documents is given to the next process of preprocess. The retrieved document preprocessing is done in module. There are two types of process is stop words removal and text stemming. Stop words are words which are filtered out prior to, or after, processing of natural language data. is the process for reducing inflected (or sometimes derived) words to their stem base or root form. Preprocess generally a written word forms. Pattern taxonomy model, that all documents are split into paragraphs. Each paragraph is considered to be each document. In each document, the set of terms are extracted. The terms, which can be extracted from set of positive documents. The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated. Inner pattern evolution used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. If partial conflict offender contains in positive documents, the reshuffle process is applied.

For instance, positive and negative statements are extracted. These term list create a unique word list, therefore easily discovered the patterns in text documents. Positive documents given to the reshuffle process. The unique combine reshuffle pattern easily identified the what documents are given to the input. In pattern deploying evaluate the term weights(supports).Here d-patterns include more semantic meaning than terms that are selected based on a term based technique. Inner pattern evolution here, only changes a patterns term support within the pattern. The normal pattern are larger pattern not supported with the huge document. In a frequent pattern of the shorten pattern fully supported in the large documents. This effective text mining method how to differ from the other documents that is specified in the following models.

III. PATTERN TAXONOMY MODEL

Assume all documents are split into various paragraphs. Set of paragraphs PS(d). D be a training set of documents that consist a set of positive D^+ and set of negative documents D^- The terms set or set of keywords are , $T = \{t_1, t_2 .... t_m\}$.

Frequent and Closed Patterns

The Given termset X in document d, covering set denote the X for d. Absolute support is the number of occurrence of X in PS(d). Relative support is the fraction of the paragraphs.

Let minimum support is the 50%. Table 1 and Table 2 illustrate the frequent patterns and covering sets.

TABLE 1

Set of Paragraphs

Paragraph	Terms
dp_1	t_1, t_2
dp_2	t_3, t_4, t_6
dp_3	t_3, t_4, t_5, t_6
dp_4	t_3, t_4, t_5, t_6
dp_5	t_1, t_2, t_6, t_7
dp_6	t_1, t_2, t_6, t_7

The various of paragraphs are splitted into following,

$$ps(d) = \{dp_1, dp_2 .... dp_6\}$$

The paragraphs contained in a various terms, terms are t_1,t_2 etc.

$$ps(d) = \{t_1, t_2 .... t_m\}$$

Term set is the subset of paragraphs in covering set. A pattern X is a closed pattern and also prove that,

$$\sup_a(X_1) \angle \sup_a(X),$$

The relative support of the covering set,

$$\sup_r(X) = \frac{[X]}{[ps(d)]}$$

The closed pattern is the

$$\sup_a(X) \angle \sup_a(X_1),$$

A term set X is called frequent pattern if its sup_r (or supa)>= min _sup, a minimum support.

If the term support are calculated based on the normal forms for all terms in d-patterns, the define the term set,

$$termset(Y) = \{t \square \forall dp \in Y \Rightarrow t \in dp\}$$

TABLE2

Frequent and Closed Patterns

Frequent Pattern	Covering Set
$\{t_3, t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_1, t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_6\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_6\}$

Frequent pattern specified by the Table 2. For example $\{t_3,t_4\}$ is always a part of the larger pattern in $\{t_3,t_4,t_6\}$. Hence shorter once is a noise pattern and except to keep the larger pattern only.

Pattern Taxonomy

Pattern be structured into a taxonomy by using is-a relation or subset relation. For example Table 1 set of paragraphs in documents. Table 1 have 10 frequent pattern and Table 2 if assuming min sub 50%.

In <t3, t4, t6>, <t1, t2>, and <t6> are closed patterns. The pattern taxonomy of the frequent pattern in table 2 represent where the nodes are frequent pattern and covering sets. non closed patterns can be pruned; the edges are "is-a" relation.

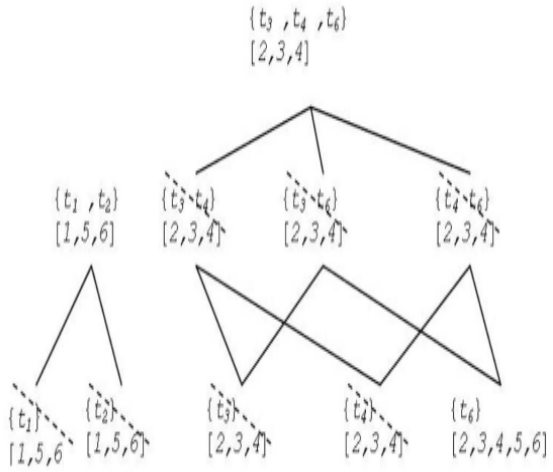

Fig 1: Pattern Taxonomy

After pruning, some direct "is-a" retaliations may changed, for example, pattern become a sub pattern of {t3, t4, t6} after pruning non closed patterns. Smaller patterns in the taxonomy, for example pattern {t6}, (see Fig. 1) are usually more general because they could be used frequently in both positive and negative documents; The semantic information used in the taxonomy and improved the performance of the closed patterns in text mining.

A Sequential pattern is an ordered list of terms. One sequence of s1 is subsequence of another sequence. Ordered Term set X in document d covering set denote the all paragraphs. Covering set specified the absolute support and relative support. Absolute support is the number of occurance of X in ps(d). The relative support is the,

$$\sup_r(X) = \frac{[X]}{\lceil PS(d) \rceil}$$

The various of closed patterns is the used to define the closed sequential patterns. A frequent sequential pattern X called closed pattern if not $\exists$ any super pattern X1 of X such that,

$$\sup_a(X_1) = \sup_a(X).$$

IV. PATTERN DEPLOYING METHOD

In text mining interpret discovered patterns by summarizing them as d-patterns in order to occurately evaluate term weights that is supports. The evaluation of term weights different to the normal term based approaches. In this pattern deploying method research, terms are weighting according to their appearance in discovered closed patterns. Here used on the xor operation.

Closed patterns represent in the pattern deploying method. Closed pattern review the composition operation. Term number of pairs p_1 and p_2. Composition of p_1 and p_2 satisfies,

$$p_1 \oplus p_2 = \{(t, x_1 + x_2) | (t, x_1) \in p_1, (t, x_2) \in p_2\} \cup \{(t, x) | (t, x) \in p_1 \cup p_2, not((t, -) \in p_1 \cap p_2)\}$$

Composition of set number pairs for example,

$$\{(t_1,1),(t_2,2),(t_3,3)\} \oplus \{(t_2,4)\} = \{(t_1,1),(t_2,6),(t_3,3)\},$$

Or

$$\{(t_1,2\%),(t_2,5\%),(t_3,9\%)\} \oplus \{(t_1,1\%),(t_2,3\%)\} = \{(t_1,3\%),(t_2,8\%),(t_3,9\%)\}$$

Weight of the terms for example, using fig 1 and Table 1, the absolute support specify the weights of the terms in the documents,

$\sup_a\left(\langle t_3, t_4, t_6\rangle\right) = 3$

$\sup_a\left(\langle t_1, t_2\rangle\right) = 3$

$\sup_a\left(\langle t_6\rangle\right) = 5$

$\hat{d} = \{(t_1,3),(t_2,3),(t_3,3),(t_4,3),(t_6,8)\}$

Term support is the total number of closed patterns that contain in the terms. Term support is the specified the weight of the terms in the documents. Frequently specify the weights of the documents in pattern deploying method. Absolute support is the closed patterns of total numbers. D-Pattern minng algorithm to improve the efficiency of the pattern taxonomy mining, also used to the Apriori property this is in order to reduce the searching space.

V. INNER PATTERN EVOLUTION

In pattern evaluation technique used to identify the noisy patterns in documents. Sometimes system falsely identified the negative documents as positive documents. So noise is occurred in positive documents. The noise patterns named as offender. If partial conflict offender contains in positive documents, reshuffle process is applied. Inner pattern evaluation how to reshuffle of terms within normal forms of d-patterns based in the training set.

The two types of offender specified that is complete conflict offender and partial conflict offender. The main functions in deploying method implement the IPEvolving algorithm. Offender of the nd is a pattern that is atleast one term in nd. Complete conflict offender removed from d-patterns in first. Then partial conflict offender, terms supports are reshuffled inorder to reduce the side effects of noise documents. Shuffling algorithm specified the xor operation in deploying method. The task of shuffling algorithm is support distribution of terms within a d-pattern. For the example the d pattern,

$\hat{d} = \{(t_1,3),(t_2,3),(t_3,3),(t_4,3),(t_6,8)\}$

Hence updated normal form by using algorithm shuffling the normal form is,

$\{(t_1,3/40),(t_2,3/20),(t_3,3/20),(t_4,3/20),(t_6,2/5)\}$

Let $\mu = 2$, offering $= \frac{1}{2}\left(\frac{3}{20}+\frac{3}{20}+\frac{2}{5}\right) = \frac{7}{20}$, and

base $= \left(\frac{3}{20}+\frac{3}{20}\right) = \frac{3}{10}$.

Hence get the following updated normal form by using the shuffling algorithm.

$\{(t_1,3/40),(t_2,3/40),(t_3,13/40),(t_4,13/40),(t_6,1/5)\}$.

Inner pattern evaluation technique useful to reduce the side effects of the noisy patterns that is low-frequency of the problem. Here changes a patterns term supports within the pattern. Threshold is usually used to classify documents into relevant or irrelevant categories. The proposed model includes two phases training phase and testing phase. In the training phase, the proposed model first calls Algorithm PTM (Dþ, min sup) to find d-patterns in positive documents (Dþ) based on a min sup, and evaluates term supports by deploying d-patterns to terms. It also calls Algorithm IPEvolving (Dþ, D_, DP, _) to revise term supports using noise negative documents in D_ based on an experimental coefficient. In the testing phase, it evaluates weights for all incoming documents. The incoming documents can be sorted on these weights.

VI. CONCLUSION

In this research work, an effective pattern discovery technique has been proposed to overcome the low frequency and problems of misinterpretation. These pattern discovery techniques include all of the association rule mining. sequential pattern mining, frequent item set mining and closed pattern mining. All frequent patterns are not specified the short pattern. In these techniques that is pattern deploying and pattern evolving proposed the effective frequency of the text in the text documents. The experimental results show that the data mining based methods and also concept based models. By comparing the other models pattern taxonomy model can improve the performance in effectiveness of the system.

VII. FUTURE ENHANCEMENT

Data mining algorithms such as association rule mining and sequential pattern mining are computationally expensive and so the pattern taxonomies-based models, especially during the phase of pattern discovery. Essential of future work an efficient algorithm of finding useful patterns from a large dataset. One possible solution to improve the efficiency of the pattern taxonomy-based model is to reduce the dimensionality of the feature space in the knowledge. The various of information lack in the selected feature, especially when the number of training examples is few. Therefore, an another one way of applying length-decreasing support constraints to frequent pattern mining. The frequent pattern may be helpful.

REFERENCES

[1] X. Zhou, Y. Li, P. Bruza, Y. Xu, R. Lau **"**A Two-stage Information Filtering Based on Rough Decision Rule and Pattern Mining" Journal of Emerging Technologies in Web Intelligence, Volume 2, No. 4, July 2010.

[2] Y. Li ,N. Zhong "Mining Ontology for Automatically Acquiring Web User Information Needs," in proc. IEEE Trans. Knowledge and Data Engineering, 2012.

[3] S. Shehata, F. Karray, M. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", in proc. IEEE Transaction Knowledge and Data Engg, 2012.

[4] Y. Xu and Y. LiGenerating, "Concise Association Rules", IEEE Transaction Knowledge and Data Engineering, 2010.

[5] A.Radhakrishnan, M. Kurian, "Effective Pattern Matching Approach for Knowledge Discovery Application", IJARECE Volume 2,Issue 2, Feb 2013

.

[6] Y. Li, J. Wu, "Summarization of Association Rules in Multi-tier Granule Mining", IEEE Intelligent Informatics Bulletin Volume 13, No.1, Dece 2012.

[7] M.Hazman, A.Rafea, "A Survey of Ontology Learning Approaches", International Journal of Computer Application, Article 6, Nov 2011.

[8] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[9] 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, pp. 2-11, 1998.

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

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

[12] [13]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.

[13]]. N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059- 1082, 2003.

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

R. Jayamari received the B.E degree in Computer Science and Engineering from the Anna University, Chennai, Tamilnadu in 2012. She is currently pursuing her M.E degree in Computer Science and Engineering with the Anna University.