International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-3, Issue - 2, March- 2015

Effective Pattern Discovery for Text Mining Using Pattern Taxonomy Model

Prof. Shriniwas Gadage, Ms. Ashwini Mandale

Abstract—

We describe an effective and innovative pattern discovery technique. In order to overcome the problem of misinterpretation and low frequency pattern taxonomy model is used. It makes use of closed sequential patterns and pruning nonclosed patterns to obtain the d-patterns. And reshuffle the terms support by using normal forms to get relevant terms from negative documents. This includes pattern deploying and pattern evolving for improving the effectiveness of using and updating discovered patterns for finding related and interested information. For deploying patterns the D-pattern mining algorithm and for evolution of patterns IPEvolving and shuffling algorithms are used. Deployment based on positive documents while evolution is based on negative documents. It requires less number of patterns for training phase. This model is effective in time complexity and coding also.

Index Terms—

closed sequential patterns, IPEvolving, PTM, term support, tfidf.

I. INTRODUCTION

To find particular patterns within an acceptable time most of the data mining techniques are proposed for the development of efficient mining algorithms. Some of techniques are sequential pattern mining, closed pattern mining and association rule mining.

Here our focus is on the development of a knowledge discovery [5] model to effectively use and update the discovered patterns and applying it to the field of text mining [1].

For finding correct features in text documents which help users to find their requirement. Some of the methods are as follows.

1) Term-based methods:
Advantages:
i) Have efficient computational performance
ii) For term weighting strong theories
Limitation:
i) There is problem of polysemy and synonymy.
2) Phrase-based methods:
Phrases have more semantic information.
Advantages:
i) Phrases are less confusing and more discriminative [3].
Limitation:
i) Terms have inferior statistical properties.
ii) Occurrence of phrases has low frequency.
iii) There are large numbers of redundant and noisy phrases.
3) Pattern mining based methods: pattern taxonomy models
Advantages:
i) Sequential patterns have good statistical properties like terms.
ii) They follow the concept of closed sequential patterns and pruned nonclosed patterns. They improve the effectiveness.

Manuscript received on 25 March, 2015
Prof.Shriniwas Gadage, Adj.faculty Computer Engg-G.H.Raisoni College of Engg and Management, Pune,India
Ms. Ashwini Mandale, ME CE student -G.H. Raisoni College of Engg and Management, Pune,India

II. LITERATURE SURVEY

In the past there are various types of text representations.

The tf*idf weighting: It uses keywords or terms as elements in the vector of the feature space. It is used for text representation in Rocchio classifiers. The Term Frequency Inverse Document Frequency (TFIDF), global IDF and entropy weighting scheme is proposed and improves performance. But the problem with this weighting scheme is how to select a limited number of features among a vast set of terms to increase the system's efficiency and avoid overfitting. To reduce the number of features, many reduction approaches have been suggested by the use of feature selection techniques. In some cases a term with higher (tf*idf) value should not have any meaning in some d-patterns. The d-patterns mean some important parts in documents.

The representation is depended on ones interest in the meaningful units of text and the natural language rules for the combination of these units. Some research works have used phrases rather than individual words with respect to the representation of the content of documents. The combination of unigram and bigrams was chosen for document indexing in text categorization [7] and evaluates on a range of feature evaluation functions. For Web document management a phrase-based text representation was proposed.

Some opinion for text representations are provided by term-based ontology mining methods [9].e.g. To determine synonymy and hyponymy relations between keywords hierarchical clustering was used. And the pattern evolution technique was introduced in order to improve the performance of term-based ontology mining.

The research works have mainly focused on developing efficient mining algorithms such as Apriori-like algorithms, Prefix Span, FP-tree, SPADE, and SLPMiner [6] for discovering patterns from a large data collection. Searching for useful and interesting patterns and rules was

an open issue. Pattern mining techniques can be used to find various text patterns, such as co-occurring terms, sequential patterns and multiple grams for building up a representation with these new types of features. Yet, the challenging issue is how to effectively deal with the large amount of discovered patterns.

K-optimal pattern technique:

It is an exploratory technique. By optimizing a user-selected objective function whilst respecting user-specified constraints it derives the k-patterns [14, 15]. It avoids the problems such as, less occurrences of most interesting patterns [16], minimum support may be irrelevant to whether a pattern is interesting, and it cannot handle dense data [17].

It allows the user to select between preference criteria and directly control the number of patterns that are discovered.

Trend analysis:

Pattern mining has been used for text databases to discover trends for text categorization (uses SPaC method), document classification and authorship identification (SVM). To describe a system for identifying trends in text documents collected over a period of time trends in text databases are used.

In authorship identification Prefix Span used to extract sequential word patterns from each sentence and used them as author's style markers in documents. The sequential word patterns are sequential patterns where item and sequence correspond to word and sentence, respectively.

III. IMPLEMENTATION DETAILS

An effective pattern discovery technique [10] is discovered. It evaluates specificities of pattern and then evaluates term weights according to the distribution of terms in the discovered patterns. It solves misinterpretation problem. It considers the influence of patterns from the negative training examples to find noisy patterns and tries to reduce their influence for the low-frequency problem. Pattern evolution is the process of updating noisy patterns evolution. As discovered patterns are more specific than whole documents the proposed approach can improve the accuracy of evaluating term weight.

PATTERN TAXONOMY MODEL

Here, we assume that all documents are split into paragraphs. Such that a given document d yields a set of paragraphs PS (d). And D is a training set of documents, which consists of a set of positive documents, D^+, and D^- a set of negative documents. Let T= {t_1, t_2, tm} be a set of terms can be extracted from the set of positive documents.

i) Frequent and Closed Patterns:

Given a termset X in document d, $\lceil X \rceil$ is used to represent the covering set of X for d, which includes all paragraphs dp ϵ PS (d) X is subset of dp i.e.

$\lceil X \rceil$ = {dp | dp ϵ PS (d)}……. (1)

The number of occurrences of X in PS (d), i.e. sup_a (X) = $\lceil X \rceil$ is called as absolute support and the relative support of pattern is present in how many fraction of the paragraphs that is, sup_r (X) = $\frac{\lceil X \rceil}{|PS(d)|}$.

Frequent pattern:

A termset X is called frequent pattern if its sup_r or sup_a ≥ min_sup, a minimum support.

Table 1 gives a set of paragraphs for a given document d, where PS (d) = {dp_1, dp_2, dp3, dp_4, dp_5, dp_6}, and duplicate terms were removed. Let min_ sup = 50%, using above definition, we can obtain ten frequent patterns in Table 1.

Table1: Set of paragraphs

Paragraphs	Terms
dp_1	t_1, t_2
dp_2	t_3, t_4, t_6
dp3	t_3, t_4,t_5,t_6
dp4	t_3, t_4,t_5,t_6
dp5	t_1, t_2,t_6,t_7
dp_6	t_1, t_2,t_6,t_7

Table 2: Frequent patterns and their covering sets.

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}

As {t_3, t_4} is short pattern it considered as a noise pattern and hence we keep the larger pattern {t_3, t_4, t_6} only

ii) Pattern Taxonomy:

By using the is-a relation structured into a taxonomy. In Table 2 we find only 3 closed patterns < t_3, t_4, t_6>, < t_1, t_2>, <t_6>.

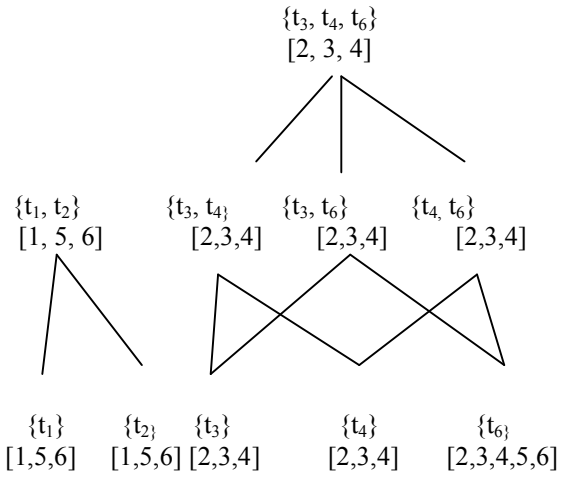

Fig 1: Pattern Taxonomy

Using Table 2 we get the pattern taxonomy for the frequent patterns as shown in Figure 1. The nodes represent frequent patterns and their covering sets. Here nonclosed patterns can be pruned. The edges represent "is-a" relation. After pruning, some direct "is-a" relations may be changed, e.g. pattern $\{t_6\}$ would become a direct sub pattern of $\{t_3, t_4, t_6\}$.

Smaller patterns in the taxonomy, e.g. pattern $\{t_6\}$, are usually more general because they could be used frequently in both relevant and irrelevant documents and larger patterns e.g. pattern $\{t_3, t_4, t_6\}$, in the taxonomy are usually more specific since they may be used only in relevant documents. To improve the performance of using closed patterns in text mining [4] the semantic information will be used in pattern taxonomy model.

iii) Closed Sequential Patterns:

If there is not exists any super pattern X1 of X such that $sup_a(X1) = sup_a(X)$ then that frequent pattern is called as closed.

b) PATTERN DEPLOYING METHOD:

We can use the semantic information in the pattern taxonomy for improving the performance of closed patterns in text mining. For this we need to understand discovered patterns by summarizing them as d-patterns means some important parts in documents in order to correctectly calculate term weights (supports). And according to their appearances terms are weighted in discovered closed patterns.

i) Closed sequential Patterns:

In order to obtain the following deployed patterns or consequential weighted patterns deploy its closed patterns on common set of terms T, for all positive documents $d_i \in D^+$,

$$\hat{d}_i = \{(t_{i1}, n_{i1}), (t_{i2}, n_{i2}), \ldots, (t_{im}, n_{im})\} \ldots\ldots\ldots\ldots\ldots (2)$$

Here tij specifies a single term and nij is specifying the total number of closed patterns that contain tij.

E.g. using Figure.1 and Table 1,

$sup_a = (<t_3, t_4, t_6>) = 3$;

$sup_a = (<t_1, t_2>) = 3$;

$sup_a = (<t_6>) = 5$;

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

Finding normal form:

Let set of d-patterns be DP in positive documents D^+, and $p \in DP$ be a d-pattern. The absolute support of term t is p (t).It is the number of patterns contains t in the corresponding patterns taxonomies. The d-patterns will be normalized in order to effectively deploy patterns in different taxonomies from the different positive documents as follows,

$$P(t) \leftarrow p(t) \times \frac{1}{\sum_{t \in T} p(t)} \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The relationship between d-patterns and terms can be described as the following,

$$\beta: DP \rightarrow 2^{T \times [0,1]} \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Such that,

$$\beta(p_i) = \{(t_1,w_1),(t_2,w_2),\ldots..,(t_k,w_k)\}, \ldots\ldots\ldots\ldots (5)$$

For all $p_i \in DP$,

Where, $\beta(p_i)$ is the normalized d-pattern of d-pattern pi.

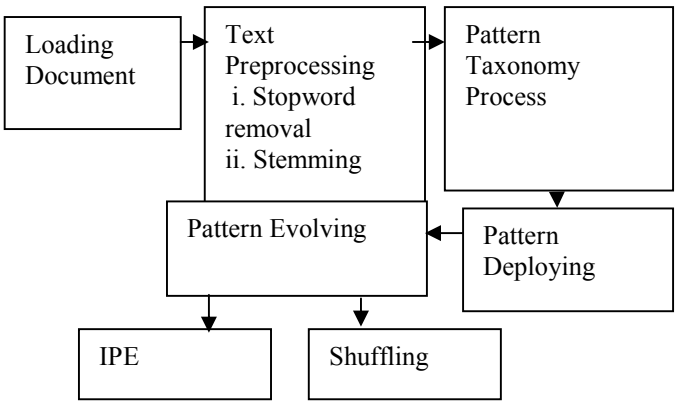

Fig 2: architecture of system

There are two phases.

Training phase:

In this the d-patterns in positive documents (D^+) based on a min_sup are found, and by deploying d-patterns to terms calculates term supports.

Testing phase:

Based on an experimental coefficient it changes term supports using noise negative documents in D^-.The arriving documents can be sorted based on these weights.

Advantages of Proposed System:

1) The proposed approach is used to improve the accuracy of evaluating term weights.
2) In this discovered patterns are more specific than whole documents.
3) It avoids the issues of phrase-based approach by using the pattern-based approach.
4) Pattern mining techniques can be used to find various text patterns.

Algorithms used:

1) D-Pattern Mining Algorithm:

For improving the efficiency of the pattern taxonomy mining, an algorithm, SPMining, is used to find all closed sequential patterns, which used the well-known Apriori property in order to reduce the searching space.

Following algorithm describes the training process of finding the set of d-patterns.
Step1: The SPMining algorithm is first called to find out the set of closed sequential patterns SP for every positive document.
Step2: All discovered patterns are composed into a d-pattern giving rise to a set of d-patterns DP.
Step3: Term supports are considered based on the normal forms for all terms in d- patterns.

2) INNER PATTERN EVOLUTION

Here we discuss how to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. This will be useful to reduce the side effects of noisy patterns because of the low-frequency problem. The technique is called inner pattern evolution, as it only changes a pattern's term supports within the pattern.

A threshold is used to classify documents into related or unrelated categories. By using the d-patterns, the threshold can be defined as follows:

$$\text{Threshold}(DP)=\min\left(\sum_{(t,w)\in\beta(p)} \text{support}(t)\right)\ldots\ldots\ldots\ldots\ldots\ldots(5)$$
$$p \in DP$$

A noise negative document (nd) in D^- is a negative document which the system falsely identified as a positive, i.e. weight (nd) ≥ Threshold (DP). To reduce the noise, we require to find which d-patterns have been used to give rise to such a fault. These patterns are called offenders of nd.

A d-pattern that has at least one term in nd is called as offender of nd. The set of offenders of nd is defined by:

$$\Delta(nd) = \{p\in DP \mid \text{termset}(p) \cap nd \neq \text{null}\}\ldots\ldots (6)$$

Two types of offenders are:
1) The offender which is a subset of nd is called complete conflict.
2) The offender which contains part of terms of nd is called a partial conflict.

The basic idea of updating patterns is, first we remove complete conflict offenders from d-patterns. In order to reduce the effects of noise documents for partial conflict offenders, their term supports are reshuffled.

The main process of inner pattern evolution is implemented by the algorithm IPEvolving,
Inputs: DP is a set of d-patterns, a training set $D = D^+ + D^-$.
Output: is a composed of d-pattern..
Step1: For finding the noise negative documents IPEvolving is used to estimate the threshold.
Step2: By using all noise negative documents revise term supports.
Step3: Discover noise documents and the related offenders.
Step4: Obtain normal forms of d-patterns.
Step5: Algorithm shuffling was called to update NDP according to noise documents.
Step6: Arrange updated normal forms collectively.

Shuffling Algorithm:

The algorithm Shuffling is to adjust the support distribution of terms within a d-pattern. For each type of offender a different strategy is dedicated in this algorithm.
Step1: Complete conflict offenders are removed as all elements within the d-patterns are held by the negative documents indicating that they can be discarded for preventing interference from these possible "noises".
Step2: The purpose of parameter offering is temporarily storing the reduced supports of some terms in a partial conflict offender. It is part of the sum of supports of terms in a d-pattern where these terms also appear in a noise document.
Step3: As termset (p) - nd≠ null calculates the base which is definitely not zero.
Step4: Updates the support distributions of terms.

IV. RESULT

As in this pattern taxonomy model approach use of stop words removal and stemming processes for text processing which will result in very effective search result. Further pattern deploying is used in which SPMining algorithm is used for finding closed patterns and D-pattern mining algorithm is used for finding d-pattern and calculates term support. Then pattern evolving uses ipevolving and shuffling algorithm to refine the discovered patterns in text documents. Figure 3 shows stemming processing.

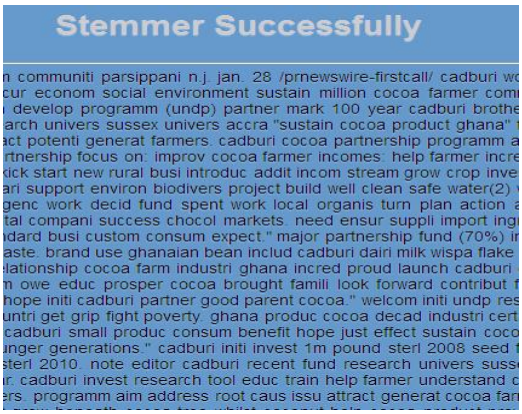

Fig 3: Stopword removal

Process starts with stopword removal then stemming is done. Then determine the number of paragraphs of document and find closed sequential patterns from each paragraph and give ID to each keyword. By listing most frequent keyword remove noise based on threshold and find most relevant document.Figure.4 shows term support evaluation.

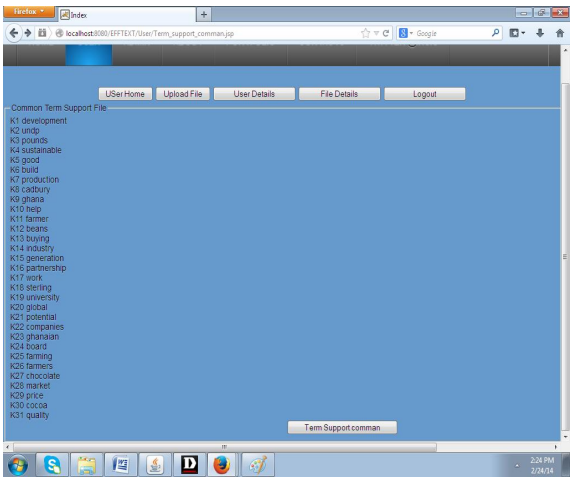

Fig 4: Term support evaluation

For evaluation of the effectiveness of pattern taxonomy model, we attempt to find the correlation between the achieved improvement and the parameter, which denotes the ratio of the number of negative documents greater than the threshold to the number of all documents. Whose value can be obtained using the following equation,

$$\text{Ratio} = \frac{|\{d|\ d \epsilon D^-, \text{weight}(d) \geq \text{threshold}(DP)\}|}{D^+ + D^-} \quad \ldots\ldots (7)$$

Comparison between PTM (IPE) and Other Models:

Figure 5 shows the number of patterns used for training, where Y axis represent patterns and X axis represent methods. By collecting the number for each topic the total number of patterns is estimated. As a result PTM (IPE) is the method that uses the fewer amounts of patterns for concept learning compared to others. This happens due to the the efficient scheme of pattern pruning is applied to the PTM (IPE) method. Even so, the classic methods adopt terms as patterns in the feature space. They use much more patterns than the proposed PTM method.

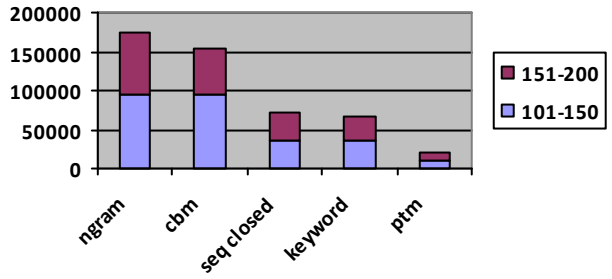

Fig 5: Comparison in the number of pattern for each model in the training method.

V. CONCLUSION

To overcome the low-frequency and misinterpretation problems an effective pattern discovery is proposed for text mining related to pattern based mining approach. The Pattern taxonomy model technique uses pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

This technique requires less time complexity as well as coding. As this is very effective model it can be useful to find data useful to lawyer, police to study the cases carefully.

ACKNOWLEDGMENT

I would like to thanks to Prof. Shriniwas Gadage, Adj.faculty Computer Engg-G.H.Raisoni College of Engg and Management, Pune for his help, Encouragement and intellectual influence, which made this paper possible. His invaluable guidance in the successful completion of this paper work.

REFERENCES

[1] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc.of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.
[2] U. Nahm and R. Mooney. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.
[3] R. Gaizauskas. An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf, 2003.
[4] M. Hearst. Untangling text data mining. In Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
[5] Y. Kodratoff. Knowledge discovery in texts: A definition and applications. Lecture Notes in Computer Science, 1609:16–29, 1999
[6] 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.
[7] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), "Experiments on Supervised Learning Algorithms for Text Categorization", International Conference, IEEE Computer Society, 1-8.
[8] Shu-Sheng Liaw, Hsiu-Mei Huang, "Information Retrieval from the World Wide Web: a User-focused Approach based on Individual Experience with Search Engines". Computers in Human Behavior, 22 (2006).
9] Fang Chen, Kesong Han and Guilin Chen (2008), "An Approach to Sentence Selection based Text Summarization", Proceedings of IEEE TENCON02, 489-493.
[10] N. Zhong, Y. Li, and S.T. Wu. Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, DOI http://doi.ieeecomputersociety.org/10.1109/TKDE.2010
[11] Han, J., Wang, J., Lu, Y., Tzvetkov, P.: Mining Top-K Frequent Closed Patterns without Minimum Support. In Int. Conf. Data Mining (2002) 211-218.
[12] Scheffer, T., Wrobel, S.: Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling. Journal of Machine Learning Research 3 (2002) 833-862.
[13] Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J.D., Yang, and C.: Finding Interesting Associations without Support Pruning. In Proceedings Int. Conf. Data Engineering, (2000) 489-499
[14] Bayardo, Jr., R.J., Agrawal, R., Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. Data Mining and Knowledge Discovery, 4 (2000) 217-240.
[15] McAullay, D., Williams, G.J., Chen, J., Jin, H.: A Delivery Framework for Health Data Mining and Analytics. Australian Computer Science Conference (2005) 381-390.
[16]Books referred: Han, J., Micheline K., 2010, Data Mining: Concepts and Techniques, San Fransisco, CA: Morgan Kaufmann publishers
[17] Website:http:\\Text mining - Wikipedia, the free encyclopedia