ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

Hybrid Approach to Improve Pattern Discovery in Text mining

Charushila Kadu[1], Praveen Bhanodia[2], Pritesh Jain[3]

Mtech(CSE) Student, Department of Computer Science and Engineering, PCST, Indore, India [1]

Head, Department of Computer Science and Engineering, PCST, Indore, India [2]

Asst. Professor, Department of Computer Science and Engineering, PCST, Indore, India [3]

Abstract: Text clustering can greatly simplify browsing large collections of documents by reorganizing them into a smaller number of patterns in text documents manageable clusters. Text clustering is mainly used for a document clustering system which clusters the set of documents based on the user typed key term. Here we proposed a hybrid model which works on reduced dimensional dataset and similarity constraints. Feature based analysis is used for reducing dimension of huge dataset. We use the feature evaluation to reduce the dimensionality of high-dimensional text vector. The system then identifies the term frequency and then those frequencies are weighted by using the inverted document frequency method. Then this weight of documents is used for clustering. Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. This model will significantly improve the result of pattern discovery in text mining.

Keywords: text mining, text classification, pattern mining, pattern deploying, pattern evolving

I. INTRODUCTION

The process of discovery of interesting knowledge in the text documents is known as text mining. Finding accurate knowledge, from the text documents to satisfy users need is still big challenge. Many term-based methods are provided by Information Retrieval (IR) such as Rocchio and probabilistic models, support vector machine (SVM), rough set models and BM25. These methods are adventitious as they perform efficiently as well as they provide very well explained theories for term weighting. But these methods, suffers from the problem of polysemy(words having multiple meanings)and synonymy (multiple words having same meaning). The semantic meaning of the word is generally confusing and make it difficult to understand what exactly user wants. For many years, many have hypothesis that phrase-based approaches are performing quiet good as they are working on semantics, but the performance is not that much encouraging. There are several reasons behind it as phrases have statically inferior properties, low frequency of occurrence and many phrases are redundant and noisy. To avoid the drawbacks of phrase-based model, new model of sequential patterns is proposed which performing better as it is statistical similarities with terms. Pattern taxonomy model have been proposed which uses closed sequential as well as pruned nonplused patterns. However, these models are not performing as per the expectations when compared with term based approaches. There are two major issues for this behavior is low frequency and misinterpretation. So, obviously it is not adequate to evaluate the weights of terms based on their presence in documents as like methods are used in IR. To solve this problem, IPE PTM is developed but it is based on large dataset. Here, we proposed new hybrid model in which pattern evolving and deployment is performed on low dimensional dataset with similarity constraints.

II. LITERATURE VIEW

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting. [1], the combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). A phrase-based text representation for Web document management was also proposed in [2].In [3]; data mining techniques have been used for text analysis by extracting co occurring terms as descriptive phrases from document collections. However, the effectiveness of the text mining systems using phrases as text representation showed no significant improvement. The likely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned in[4]. In, hierarchical clustering [5], [6]was used to determine synonymy and hyponymy relations between keywords. Pattern mining has been extensively

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

studied in data mining communities for many years. These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [7], [8], [9]. In the field of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent item sets, co-occurring terms and multiple grams, for building up a representation with these new types of features. Nevertheless, the challenging issue is how to effectively deal with the large amount of discovered patterns. 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. Pattern taxonomy model was also developed in [11] and [10] to improve the effectiveness by effectively using closed patterns in text mining. In addition, a two-stage model that used both term-based methods and pattern based methods was introduced in [12] to significantly improve the performance of information filtering. Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model [13], [14] was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. pattern based methods was introduced in [12] to significantly improve the performance of information filtering.

A. Pattern Taxonomy Model

The basic definition of sequences used in this study is described as follows. Let $T = \{t1, t2, tk\}$ be a set of all terms, which can be viewed as keywords in text datasets. A sequence $S = <s1, s2\ldots sn>$ ($si \in T$) is an ordered list of terms. A sequence $\alpha = <a1, a2,\ldots, an>$ is a subsequence of another sequence $\beta = <b1, b2,\ldots, bm>$, denoted by $\alpha \subseteq \beta$, if there exist integers $1 \leq i1 < i2 < \ldots < in \leq m$, such that $a1 = bi1$, $a2=bi2,\ldots$, $an = bin$. The sequence α is a *proper* sub-sequence of β if $\alpha \subseteq \beta$ but $\alpha \neq \beta$, denoted by $\alpha \subset \beta$. 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*.

III. PROBLEM STATEMENT

There are many methods described in information retrieval for text mining but they are suffering from many disadvantages such as hyponymy and polygamy. In effective pattern discovery model clustering is performed on high dimensional set which is basically large set of documents. Though existing model is using pattern evolving and deploying methods which are efficient but it is not considering similarity relationship of the words, as the other semantic algorithm faces problems like polygamy and other. As semantic and similarity analysis of the documents is not considered for clustering, results are not significantly improved. Secondly, existing model uses directly high dimensional set for clustering in terms of its frequency, hence, there is probability of grouping of higher frequency contents together into one cluster even thought they are unrelated to each. As a result this is not an effective technique for pattern discovery in text mining for low frequency terms. So to handle the low frequency term, we are trying to make some changes to the existing algorithm. So that it can be used in hybrid manner for both high dimensional as well as low dimensional dataset.

IV. PROPOSED SOLUTION

A Hybrid model for cluster based pattern discovery for low dimension set in text mining combines the best features of different pattern discovery algorithm in text mining. This proposed model works in various phases which are as follows.

A. Data preprocessing
B. Semantic based analysis
C. Similarity-based analysis
D. Pattern evolving and Pattern mining

Following is the description of each phase of the model.

A. Data Preprocessing

In data preprocessing, we use the feature evaluation to reduce the dimensionality of high-dimensional text vector. The system then identifies the term frequency and then those frequencies are weighted by using the inverted document frequency method. Then this weight of documents is used for clustering. Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification. Our proposed works presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving.

B. Semantic based Analysis

In this phase, documents are divided into the groups based on semantic analysis. Here, document classification is performed on the meaning of the word. But in this phase there are problem of having one word with different meaning and different word with same meaning. So, in this phase classified set of data is not perfectly accurate. It may lead to the mix type of documents in one cluster.

C. Similarity Based Analysis

Output of the second phase will be input to this phase. Here, semantic based classification will further go in analysis phase which is based on similarity analysis. The output of this phase will result into similar types of documents in one cluster.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

D. Pattern Evolving and mining

In this phase desired pattern are evolved from the clusters obtained from above phases. This is important phase of the model which actually evolves patterns which will match to the keywords of user who want relevant information from large database which are generally in electronic forms.

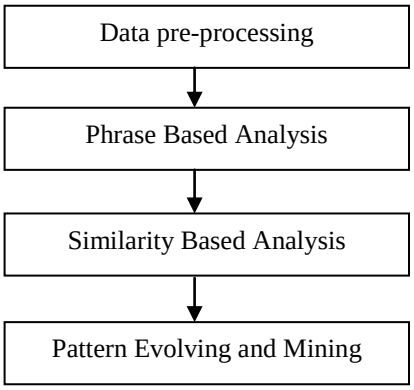

Fig. 1. Phases in Hybrid model for cluster based pattern discovery in low dimension set in text mining

V. ARCHITECTURE OF HYBRID MODEL

S This a architecture supports best feature of PTM and IPE methods of pattern mining along with similarity constraints which is not considers in existing model.. This is 3-tier architecture in which operational are performed on different input in order to find out exact pattern which is matching to the need of user. In first tier, documents are converted from high dimensional form into low dimensional one. Once we get the small dimensional set of documents, clustering is performed to get the separate set of positive and negative documents. Pattern taxonomy model is used for generating patterns. But the main problems remain how to select, update and deploy the pattern which is exactly required by the user. These patterns are then evaluated on the basis of semantic analysis. Finally output is considers on similarity basis.

Hybrid model for pattern discovery for low frequency set as well as high frequency set in text mining combines the best features of different pattern discovery algorithm in text mining. This proposed model works in various phases which are as follows.

- SP Mining.
- PTM
- Inner Pattern Evolving

A. SP Mining

In this module we generate a frequent sequential pattern is a maximal sequential pattern if there exists no frequent sequential pattern. The length of sequential pattern indicates the number of words (or terms) contained in pattern. A sequential pattern which contains *n* terms extracted from given set of documents. Here we take set of documents as input we generate term sequences. In this module we present a pattern-based model PTM (Pattern Taxonomy Model) for the representation of text documents. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. Once the tree is constructed, we can easily find the relationship between patterns. The next step is to prune the meaningless patterns in the pattern taxonomy.

B. PTM

In this module we take positive documents and negative documents and we adjust the term weights based on term weight of positive document and negative document. Using this technique we can increase maximum likelihood event one documents having more overlapping terms and less content of the document we get accurate results.

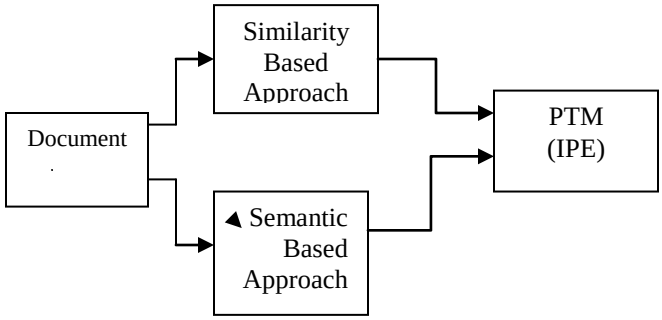

Fig. 2. Architecture of Hybrid Model

For this model, searching term is in put for this hybrid model. Results are generated in terms of relevant documents.

VI. EXPERIMENTAL RESULT

All In this study, Reuter's text collection is used to evaluate the proposed approach. Term stemming and stop word removal techniques are used in the prior stage of text preprocessing. Several common measures are then applied for performance evaluation and our results are compared with the state-of-art approaches in data mining, concept-based, and term-based methods.

For this study, I have used Reuter's text Collection to evaluate proposed hybrid model. Natural Language processing techniques like term stemming and stop ward removal is used in the prelims steps. For evaluating performance, several common measures are applied such as b/p known as break point where precision is equal to recall, interpolated average precision and median average precision and result of proposed hybrid model are compared with the several state-of-art approaches in the field of text mining like concept base model and term base model.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

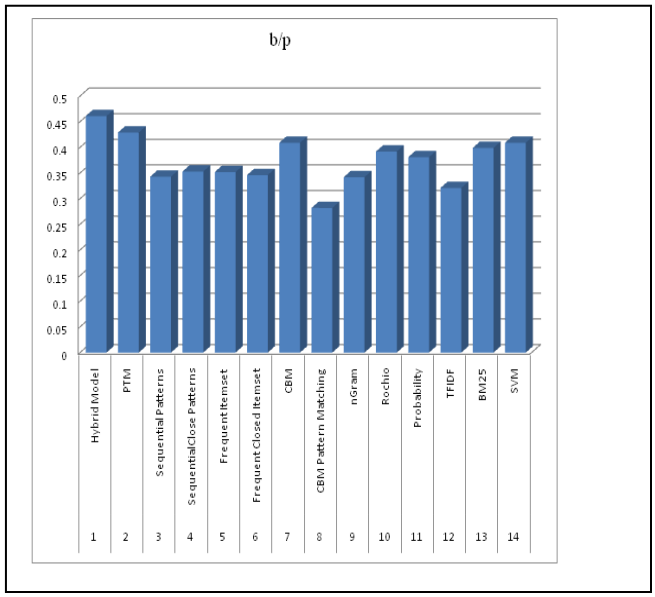

Fig. 3. Comparison of IR Methods on the basis of B/P

From the above graph, the most important information we get is that this proposed 3-tier hybrid model is performing well when compared with pattern mining approaches, term-based methods as well as concept-based methods.
Significant differences are not measured on time complexity when this model is compared with all other model used in text mining.

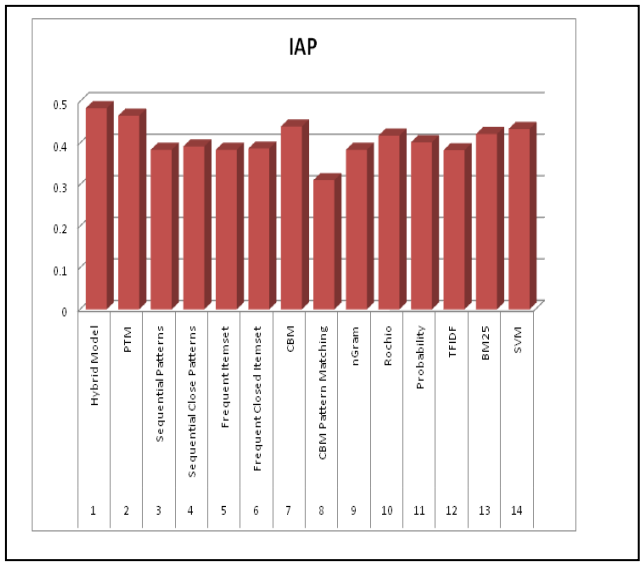

Fig. 4. Comparison of IR Methods on the basis of IAP

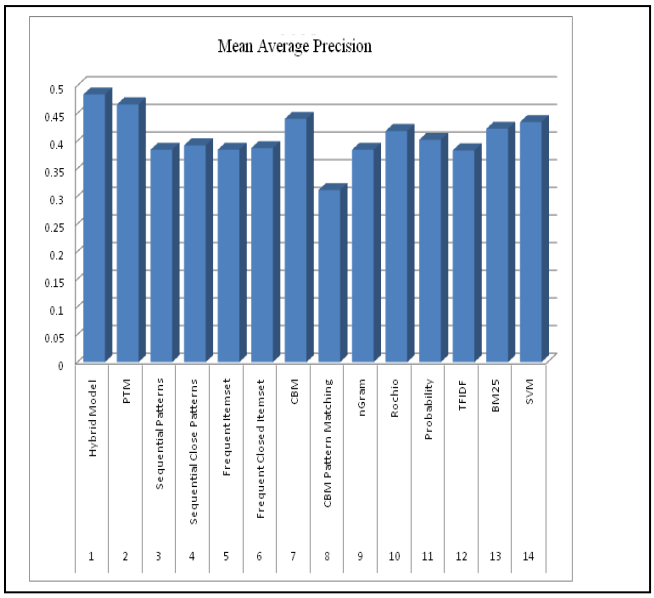

Fig. 5. Comparison of IR Methods on the basis of IAP

CONCLUSION

When we analyses the result of this model, with the previous model which are used in text mining, we come to know that this model is performing quite well. As we combine both the semantic and similarity constraints, in this model, we are able to overcome the problems occurring in both approaches. Also, in this model we have combined the PTM (IPE); hence the results are significantly improved.
In future, we can personalize this model, so that results or data extracted from the text document is related to the user profile. Here we will, generate the user profile for the system who is searching the information. Then we will extract the documents which are matching to his profile.

ACKNOWLEDGMENT

I expressed my heartfelt thanks to my guide Prof. Pritesh Jain for his continuous guidance and encouragement throughout this paper. I am also sincerely thankful to my HOD Prof. Praveen Bhanodia, for his continuous support and guidance. Finally, I am thankful to GOD and my family for always being with me.

REFERENCES

[1] S. 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.

[2] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo,"Applying Data Mining Techniques for Descriptive PhraseExtraction in Digital Document Collections," Proc. IEEE Int 'lForum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

[3] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM
[4] SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), pp. 37-50, 1992.
[5] Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.
[6] Manning and H. Schu¨ tze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.
[7] 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.
[8] Y. Li and N. Zhong, "Interpretations of Association Rules by Granular Computing," Proc. IEEE Third Int'l Conf. Data Mining
[9] (ICDM '03), pp. 593-596, 2003.
[10] Y. Xu and Y. Li, "Generating Concise Association Rules," Proc.ACM 16th Conf. Information and Knowledge Management (CIKM '07), pp. 781-790, 2007.
[11] 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
[12] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data
[13] 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
[14] 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.
[15] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval, vol. 1, pp. 69-90, 1999.
[16] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99), pp. 42-49, 1999.

.