Pattern Taxonomy Term Based Model for Text Document Classification

[1]S.Brindha, [2]Dr.S.Sukumaran
[1]Ph.D Scholar, [2]Associate Professor,
Department of Computer Science, Erode Arts and Science College, Erode,Tamilnadu, India

Abstract: **The quality of discovered related features in text documents are describing based on user preferences. For the reason that of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. Most of the problems are occurred in polysemy and synonmy. Over the years, there has been repeatedly held the hypothesis that pattern-based methods should achieve better than term-based ones. Big challenge is how to effectively use large scale patterns vestiges a hard problem in text mining. In this paper, the robustness is used to discuss the characteristics of a model for describing its training sets is distorted or the application environment is altered. A new model robust if it still provides satisfactory performance regardless of having its training sets are altered or changed. To make a breakthrough in this challenging issue, this paper presents a pioneering model for weight feature discovery. It discovers both positive and negative patterns in text documents as at a higher level features and deploy them over low-level features. The terms also classify into categories and updates term weights depends on their specificity and their distributions in patterns. Significant experiments using this model on RCV1, TREC topics and Reuters-21578 significant experiments using this model on RCV1, TREC topics and Reuters-21578 demonstrate that the proposed model significantly outperforms both the state of the Pattern Taxonomy Term-based methods and the pattern based methods.**

Keywords: **Classification, Pattern Mining, Term Frequency, Weights, Pattern Taxonomy, TF-IDF.**

I. Introduction

With the hasty expansion of the World Wide Web, the mass of online text data has grown at very fast speed in recent years. The objective of weight document discovery of text classification [14] is to find the useful features available in text documents, as well as both applicable and inappropriate ones, for describing text mining results and find the weights of the text document. This is a particularly challenging task in modern information analysis, from both an empirical and theoretical perspective. This problem is also of central interest in many web modified applications, and it has acknowledged consideration from researchers in empirical and theoretical perspective. This problem is also of central interest in most of the web personalized applications, and it has received more attention from researchers in Data Mining, Machine Learning, and information Retrieval [10] and Web Intelligence communities. There are two challenging issues in using pattern mining techniques for discovery relevance features in both related and unrelated documents. The first is the low-support problem. Given a topic, long patterns are usually appearing more specific for the topic, but they usually appear in documents with maintain or regularity. If the smallest amount sustain is decreased, a lot of noisy patterns can be discovered. The second problem is the misinterpretation problem, which revenue the measures used in pattern mining [12] turns out to be not suitable in using patterns for solving harms. For model, a greatly frequent pattern may be a general pattern since it can be frequently used in both related and inappropriate documents. There are several existing methods for solving the two searches out issues within text mining. Pattern Taxonomy mining models [11] have been proposed, in which mining closed sequential patterns [15] in text paragraphs and deploying them over a term space to weight useful features. Concept-based model (CBM) has in addition been proposed to realize concepts by using ordinary language processing (NLP) techniques. These patterns based or concepts based approaches have shown an important improvement in the effectiveness. However, fewer significant improvements are made compared with the best term based methods because how to effectively integrate patterns in both relevant and irrelevant documents is still an open difficulty. Over the years, people have residential many full-grown term-based techniques for ranking documents, information filtering and text classification. Recently, several hybrid approaches were proposed for text classification.

Term selectivity determines the scope to which the term focuses on the topic that user's requirements. It is very complicated to compute the precision of terms because a term's specificity depends on user's perspectives of their information needs. The first definition of the specificity of terms because a term's specificity which calculated the score of a term based on its appearance in discovered positive and negative patterns. In agreement with the diffusion of terms in a training set, it implement a new definition for the specificity function and used two experimental parameters to cluster terms into three categories: "positive", "general', and "negative". The RFD model is able to precisely appraise term weights according to equally their specificity where the higher level features include both positive and negative patterns. The term classification method proposed in requires manually setting two empirical parameters according to hard sets. The RFD model and empirically demonstrate that the proposed specificity function is logical and the term classification can be successfully approximated by a feature clustering method. Design an inclusive approach for evaluating the text document models. In addition, conducted some new experiments by using six new descending windows to adaptively update the training sets and also applying the RFD model for dual text classification to test the

toughness. Used for this evaluation, only use RCV1 because Reuters-21578's [9] difficult set will become too small if we increase training sets.

II. Literature Review

The literature consists of effective work done for the various mining techniques available which are as follows.
G. Chandrashekarand F. Sahin, addresses that A survey on feature selection methods term based method suffers from the problems of polysemy and synonmy and they suggested that Pattern based [7] method performs better than term based methods. For finding relevant information they have used processes of Pattern Deploying and Pattern Evolving. Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S.Kamel, Fellow, IEEE proposed concept based mining technique which calculates similarities in documents by matching concepts between documents and also by semantics of sentences. R. Bekkerman and M. Gavish, High-precision phrase-based document classification on a modern scale [5], used the concept of association rule to evaluate differences between Positive and negative patterns in the documents. They have proposed Association rule mining algorithm to extract useful patterns from the large data. Novovicova et al. used SFS that took into account, not only the mutual information between a class and a word but also between a class and two words [18]. The support vector machine technique is a popular and highly accurate machine learning method for classification problems was introduced in the early 1990s[22]. In 1998, the study proposed by Joachims explored the benefits of using SVM for text Categorization [21]. The results were slightly better. Lim proposed a method which improves performance of KNN based text classification by using well estimated parameters [14]. Some variants of the KNN method with different decision functions, k values, and feature sets were proposed and evaluated to find out adequate parameters.

III. Theory Discussion

A document classification is described graph-based approach. The representation of graph offers the advantage that it allows for a to a great extent additional significant document encoding than the more standard bag of words/phrases approach, and accordingly gives better classification precision. Document sets are represented as graph sets to which a weighted graph mining algorithm is functional to take out frequent sub graphs, which are then further processed to produce feature vectors for classification. Weighted sub graph mining is used to ensure classification effectiveness and computational efficiency. The most significant sub graphs are extracted. In a real world textual data set pattern Taxonomy approach is validated using several popular classification algorithms together. Some of the dataset used this text classification algorithm to give a performing result. When the size of dataset increased, further processing on extracted frequent features is essential.

The most common document formalization for text classification is the vector space model identified on the bag of words/phrases demonstration. The main advantage of the vector space model is that it can readily be employed by classification algorithms. The bag of words/phrases representation is suited to capturing only words/phrases frequency structural and semantic information is ignored. It have been recognized that structural information plays an important role in classification accuracy. An alternative to the bag of words/phrases representation is a graph based illustration, which spontaneously process much more communicative power. However, this illustration introduces an additional level of complication with the intention of the computation of the similarity between two graphs is significantly more computationally expensive. Some work has been done on hybrid representations to capture both structural elements and significant features using the vector model.

An approach to text classification using a graph based representation has been specified. The graph demonstration of text allows adding both the structure and content of documents to be represented. Key constructs to sustain text classification can then be identified using frequent sub graph mining. The disadvantage of normal frequent sub graph mining is that it is arithmetically expensive, to the extent that any potential advantage of the graph representation of text cannot be realized. To overcome this disadvantage a weighted subgraph mining mechanism is proposed, w-g Span. In effect W-g Span selects the most significant constructs from the graph representation and utilizes these constructs as input for classification. Investigational evaluation demonstrates that the technique works well, significantly out-performing the unweighted approach in each and every case. A number of diverse weighting schemes were considered coupled with three different categories of classifier originator. In terms of the generated classification accuracy weighting outperformed the other proposed weighting mechanisms. Weighting also worked well in terms of computational efficiency and to represents the greatest overall weighting strategies.

To describe related features for a given topic, normally believe that specific terms are very useful in order to distinguish the topic from other topics. However, our experiments show that using only specific terms is not good quality sufficient to improve the performance of relevance feature discovery because user information [20] needs cannot simply be covered by documents that contain only the specific terms. Based on theory discussion this issue in the evaluation section. This section discusses the testing environment, and reports the experimental results and the discussions. The use of specific terms and general terms for recitation user information needs. Supervised approach that needs a training set together with both significant documents and irrelevant documents. Most of the users used two accepted data sets to test the proposed model: Reuters Corpus Volume 1 is one of the very large data collections; and Reuters-21578 [9], a small one. There are 806,791 documents including RCV1 that covers a broad spectrum of issues or topics.

Drawbacks

The challenging issue for text feature selection [1] in text documents is the recognition of which format and also fined where the related features are in a text document. The improved effectiveness was not significant. Building an information filtering model that matches user needs to user profiles is a complex challenge. They had low frequency patterns [4], the high-level patterns are deployed into low-level terms.

The proposed model is called the relevance feature discovery model, and consists of three major steps: feature discovery deploying, term classification and term weighting. First finds positive and negative patterns and terms in the training set. Finally, Pattern Taxonomy methods, mechanism out the term weights by utilizing Algorithm WFeature.

IV. Proposed Methodology

Pattern Taxonomy Term Based Models (PTTM) that utilize closed sequential patterns [15] in text documents to overcome the limitation of traditional term-based approaches. However, the key challenge of PTM is how to effectively deal with numerous discovered patterns for the extraction of accurate features. Among discovered patterns, there are many meaningless patterns, and also some discovered patterns may include general information (i.e., terms or phrases) about the user's topic. Such patterns are noisy and often restrict effectiveness. This chapter presents a novel data mining framework for acquiring user information [2] needs or preferences in text documents. This framework utilizes pattern taxonomy mining to capture important semantics information in a feedback set of relevant documents. After that, a new post-mining method, named pattern cleaning, for relevance feature discovery, is applied to reduce the effects of noisy information captured by pattern mining. A Set of Textual Documents are arranged likewise dc1,dc2,dc3….etc., and the words of the Documents Words wd2,wd3…etc.

dc1 -- wd2 wd3
dc2 -- wd4 wd5 wd6
dc3 -- wd4 wd5 wd6 wd7
dc4 -- wd4 wd5 wd6 wd7
dc5 -- wd2 wd3 wd6 wd7
dc6 -- wd2 wd3 wd6 wd7
dc7 -- wd1 wd2 wd3 wd4

For a given topic, closed sequential patterns extracted from positive (relevant) documents (or positive patterns for short) capture prices of meaningful information for describing user information needs. Frequent Words are described here like wd4, wd5, wd6 etc., Covering set of the particular document as dc2, dc3, dc4 etc.,

{wd4,wd5,wd6} {dc2,dc3,dc4}
{wd4,wd5} {dc2,dc3,dc4}
{wd4,wd6} {dc2,dc3,dc4}
{wd4} {dc2,dc3,dc4,dc7}
{wd6} {dc2,dc3,dc4,dc5,dc6}
{wd2,wd3} {dc1,dc5,dc6,dc7}
{wd2} {dc1,dc5,dc6,dc7}
{wd3} {dc1,dc5,dc6,dc7}

In figure.1 Pattern Taxonomy Processing method shows process starts with input documents. Text documents are in XML format as well as Word document or PDF document. System uses various data collected from different field which is collected and used in training set and test set. Documents are processed in positive and negative. Positive means document is relevant to the topic. The irrelevant documents are negative. System uses closed frequent patterns as well as non sequential closed pattern for finding concept from dataset. Each document is divided into paragraph. Those documents are independently processed using pre-processing step. These steps include the collection of the document preprocessed in stop word removal and text stemming, and finally calculate the term weight calculation. Dimension reduction process in done to get feature set with low support dimension reduction to get the feature set. Sequential pattern and closed sequential pattern is discovered by pattern discovery model.

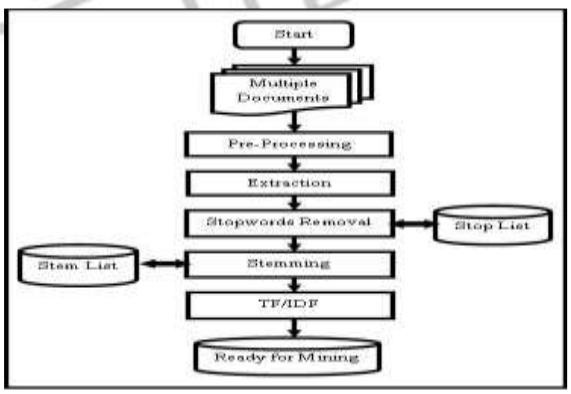

Figure.1 Pattern Taxonomy Preprocessing Method

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words is critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. Stop words are generally thought to be a "single set of words". It really can mean different things to different applications.

- Determiners - Determiners tend to mark nouns where a determiner usually will be followed by a noun examples: *the*, *a*, *an*, *another*

- Coordinating conjunctions – Coordinating conjunctions connect words, phrases, and clauses examples: *for, an, nor, but, or, yet, so*
- Prepositions - Prepositions express temporal or spatial relations examples: in, under, towards, before.

Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents.

Stop words are very frequently used common words like 'and', 'are', 'this' etc. They are not useful in classification of documents. So they must be removed. However, the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the system performance. Every text document deals with these words which are not necessary for text mining applications. Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pronouns, etc. that does not give the meaning of the documents. These words are treated as stop words.

Stemming is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words: "presentation", "presented", "presenting" could all be reduced to a common representation "present". This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented.

However, a lot of meaningless and irrelevant information available in the feedback documents can easily affect the quality of extracted features. Using closed pattern mining cannot deal with the noisy information subject to a specific need of user. For example, short patterns with high support generally contain general information for a specified topic, but specific long patterns have low support. The objective of pattern cleaning is to reduce the effects of noises caused by the discovery process. The main idea of pattern cleaning is to utilize non-relevant information to refine the relevant knowledge for a specified topic. However, using all negative documents may be not interesting and increase noises since they maybe often collected from other topics. In this research, introduced the notion of offenders to address the above issue. An offender is defined as a negative document that is closer to positive ones.

Mining useful features to help users searching for relevant information is a challenging task in information retrieval and data mining. User relevance feedback is the most valuable source of information to acquire information needs of individual users. However, too much noise available in real-world feedback data can adversely affect the quality of extracted features. The major research issue in this method is how to extract useful knowledge in user relevance feedback to reduce the effect of weight features extracted by frequent pattern mining. It also presents a new pattern-based approach to relevance feature discovery. Introduce the concept of pattern cleaning, refining the quality of discovered frequent patterns in related documents using the selected non-related samples. Show that the information from the non-significant samples is very useful to reduce noisy information in significant documents as well as improve the quality of specific features to retrieve accurate information.

V. Deploying Higher Level Patterns

In support of term-based approaches, the usefulness of a weighting, particular term is based on its manifestation in documents. On the other hand, for pattern-based approaches [15], weighting the usefulness of a given term is based on its appearance in discovered patterns [11][12]. To get better the effectiveness of the pattern taxonomy mining, an algorithm, SP Mining (D+, ;min_sup), was proposed to find closed sequential patterns [15] for all documents ∈D+, that is utilized to the well-known Apriori property to diminish the penetrating space. For all applicable documents di ∈ D+, the SP Mining algorithm discovers all closed sequential patterns, SPi based on a specified min_sup. Processing of discovered patterns is carried. The discovered patterns are organized in specific format using Pattern Deploying method (PDM) Algorithms. It is mainly organized discovered patterns in term, frequency form and also combining discovered pattern vectors. Pattern Deploying with Support gives same output as PDM with support of each term. Discovered patterns are deployed methods, and then the pattern evolving process is used to refine patterns. It representing the concept of the topic is generated eventually. Each document in the test set is accepted by the Test module and the related documents to topic as an output. The result of data transform is a set of transactions and each transaction considered as a vector of stemmed terms. By splitting each and every document into several transactions are used to find the frequent patterns from the textual documents. The estimation of every document in the test dataset is conducted using the document evaluating Test process. After testing the document performance calculated using the metric such as precision and recall and f1 measures.

VI.Pattern Taxonomy Term Based Method

The term graph model is an improved version of the vector space model [13] by weighting each term according to its relative "importance" with regard to term associations. Specifically, for a text document Di, it is represented as a vector of term weights Di =< w1i, . . . , w|T|i >, where T is the ordered set of terms that occur at least once in at least one document in the collection. Each weight wji represents how much the corresponding term tj contribute to the semantics of document di. Although a number of weighting schemes have been proposed (e.g., boolean weighting, frequency weighting, tf-idf weighting, etc.), those schemes determine the weight of each term individually. As a result, important yet rich information regarding the relationships among the terms are not captured in those weighting schemes. We propose to determine the weight of each term in a document collection by constructing a term graph. The basic steps are as follows:

1. Prepocessing Step: For a collection of document, extract all the terms.

2. Graph Building Step:
(a) For each document, we view it as a transaction: the document ID is the corresponding transaction ID; the terms contained in the document are the items contained in the corresponding transaction. Association rule mining algorithms can thus be applied to mine the frequently co-occurring terms that occur more than minsup times in the collection.
(b) The frequent co-occurring terms are mapped to a weighted and directed graph, i.e., the term graph.

$$IDF\ (t,d)= \frac{|D|}{\text{no of Document t appears}}$$

Then Term Frequency - Inverse document frequency [TF-IDF] is calculated for each word using the formula,

$$Tfidf(t,f,d)=tf=(t,d) * idf\ (t,d)$$

In above equation tf, d denotes the frequency of the occurrence of term t in document d. TF-IDF is calculated for each term in the document by using Term Frequency (Tft, d) and Inverse Document Frequency (idft, d).

VII. Weighting Features

The calculation of unique RFD term weighting function includes two steps: initial weight calculation and weight revision. In this paper integrate the four steps into the following equation:

$Wd(t)=\{dc\_sup(t,D^{+})(1+spe(t))$ t€T^{+}
$Wd(t)=\{dc\_sup(t,D^{+})$ t€G
$Wd(t)=\{dc\_sup(t,D^{+})(1-|spe(t)|)$ t€T_1
$Wd(t)=\{-dc\_sup(t,D^{-})(1+|spe(t)|)$

Documents in both RCV1 and Reuters-21578 [9] are described in XML. To keep away from bias in experiments, all of the information regarding the meta-data was unobserved. Every document were treated as plain text documents by a preprocessing, including removing stop-words according to a certain stop-words list and stemming terms by applying the Porter Stemming algorithm. The weighting features are move to original RFD term weighting function. The initial weight calculation should be calculated from RFD weighting function. Finally weight revision calculated.

Term classification is RFD uses both specific features [3] (e.g., T+ and T_) and general features (e.g.G). Therefore, the key research question is how to find the best partition (T+, G , T_) to successfully classify relevant documents and irrelevant documents. For a given set of features, however, this question is an N-P solid problem because of the bulky number of possible combinations of groups of features. In this section suggested an estimate approach, and efficient algorithms to refine the RFD model. In term-based approaches, the evaluation of term weights (supports) is based on the distribution of terms in documents. The evaluation of term weights (supports) is different to the normal term-based approaches. PTM is implemented by three main steps: 1) discovering useful patterns [19] by integrating sequential closed pattern mining algorithm and pruning scheme; 2) using discovered pattern deploying; 3) accommodate user profiles by applying pattern evolution.

Pattern Taxonomy Weighting Algorithm

Train
Step:1 Taking positive and negative documents to train
Step:2
Start
B1= positive document
B2 = negative document
If (B1== Tech) //* Tech = Technology document
{
Print("click on positive document to choose"); }
Else
{
If(B2==Sci) //* Sci=Science document
Print("click on negative document to choose");
}
//Select folder for negative document to train
Step:3
Step:4
If(b1==b2)
Print("select different path to choose positive and negative documents");
else
//perform Pattern Taxonomy Method to find out frequent pattern and weights of the individual patterns as well as terms
Test
Step5: Testing accuracy among the documents which have choose
Step6: Select both documents to compare accuracy

```
if (b1==b2)
        Result is false positive since it has to be selected positive
        and negative documents for inputs
else
        result is true positive and result is accurate
Prediction
Step7: Take a document as input
if
        Taken file is related to technology it classifies the
        document as positive document
else
        It is negative document
End
```

Finally testing phase finds differences in positive/negative documents by the centroid obtained in training phase by ranking each of them. The easy method to approximate similarity between documents and centroid by summing weights of patterns which are in the documents. Each and every document is preprocessed with word stemming and words removal in to a set of transactions based on its nature of document structure. System selects one of pattern taxonomy algorithm to extract the pattern based documents.

VIII. Experimental Results

To determine accurate measures of similarity or difference between documents depict results by graph pattern and table pattern. The experimental setup consists of relevant documents that are termed as positive and negative documents .i.e Technology related (Positive) and Science related (Negative). That take into account support factor also consisting of support = 50, support = 75 and support= 90. And also take training set of 6 documents in which take 3 positive documents and 3 negative documents. Basing on the support factors have calculated documents weights and also uniquely calculated document weights by PTM method which is discussed earlier.

Doc/Sup	50	75	90
Doc 1	1.31556	1.80357	1.80357
Doc 2	1.75182	1.87012	1.87012
Doc 3	2.13338	2.27178	2.27178
Doc 4	1.56941	1.0743	1.0867
Doc 5	1.51786	1.9243	1.8243

Table1. Support Weight Factors for Diverse Documents

In the above Table.1 have 5 documents with their relevant weights basing on the support factor. The collection of document may be in the order of 50, 75, and 90. Compute the above table in pictorial pattern which helps in easy way of understanding. It has to be continued to develop the RFD model and experimentally prove that the proposed specificity function is realistic and the term classification can be effectively approximated by a feature clustering method.

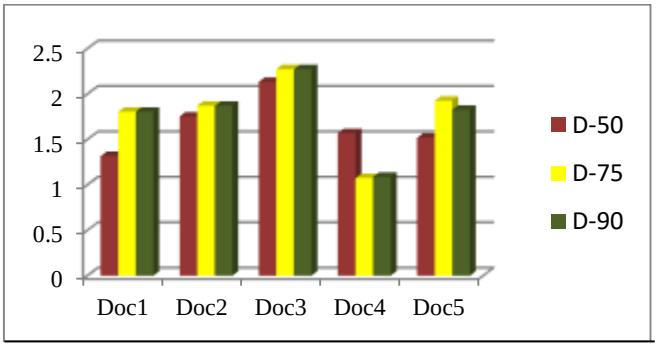

Figure.2 Patterned Document Support Weight Frequency

The above chart is also displayed the document to be patterned using support weight frequency. Considered five different documents to be compared using Relevance feature discovery method. The comparison of documents can be analyzed using patterned text classification methods. Based on the weight features and noisy features the results should be varied for different document. Finally they receive an expected performance, but it need requires the manually testing a large number of different values of parameters. After Test process, the system is evaluated using three performance metrics precision, recall and F1-measure.Using these metrics, different methods are compared to check the most appropriate method which gives maximum relevant documents to topic. Reuters-21578 dataset consist of 90 topics. Comparison of precision, recall and f1-measure for topic ship by considering top-k documents with highest relevance score.

Terms	Patterns
1-term	Pct Offer River Ship Strike Seaman Sector Redund
2-term	Offer pct Offer river Strike pai Pai seamen Strike seamen
3-term	Pct river offer pai strike seamen ship sourc capac industri ship japan`

Table 2: 1-term, 2-term, 3-term patterns

Document No.	Term	Support
23	River	1.0
43	Ship	0.25
54	Seamen	0.25
62	Missil	0.25
63	Yard	1.0
81	Industry	0.25
62	Sourc	0.25
98	Shell	1.0
98	Strike	1.0
128	Protect	1.0

Table.3. Patterns After Pattern Evolving

TP (True positives) is the number of documents the system correctly identifies as positives; FP (False Positives) is the number of documents the system falsely identifies as positives; FN (False Negatives) is the number of relevant documents the system fails to identify. The precision of first K returned documents top-K is calculated. The precision of top K returned documents refers to the relative value of relevant documents in the first K returned documents.

IX. Scope of Feature Work

The proposed method achieves the best performance for comparing with term-based baseline models and pattern-based baseline models. The consequences also show that the term classification can be effectively approximated by the proposed clustering method. Compared with the primary model, the novel representation is to a great extent additional capable and achieved the satisfactory performance as well. In this paper also includes a set of experiments. To Increase weight values and consider huge number of documents and compared those documents using various Pattern algorithm.

X. Conclusion

The main aim of research an approach should be in alternative for relevance feature discovery in text related documents. Using RFD method to find the low-level features and also classify the text based on both their appearances in the higher-level patterns and their specificity. Also introduces a method to select non related documents for weighting features. The first Relevance Feature Discovery model uses two experimental parameters to set the boundary between the categories. The new method uses a feature clustering technique to automatically group terms into the three categories. The proposed methodology is

reasonable and robust. This paper demonstrates the new models totally tested and prove the results statistically significant. The paper also proves that the use of unrelated opinion is considerable for improving the performance of relevance feature discovery models. A promising methodology for developing effective text mining models for RFD discovery based on both positive and negative document.

References

[1] Aghdam.M N. Ghasem-Aghaee, and M. Basiri, "Text feature selection using ant colony optimization," in Expert Syst. Appl.,vol. 36, Pp: 6843–6853, 2009.
[2] A. Algarni and Y. Li, "Mining specific features for acquiring user information needs," in Proc. Pacific Asia Knowl. Discovery DataMining, Pp: 532–543, 2013.
[3] A. Algarni, Y. Li, and Y. Xu, "Selected new training documents to update user profile," inProc. Int. Conf. Inf. Knowl. Manage.,Pp: 799–808, 2010.
[4] Azam N and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization,"Expert Syst. Appl., vol. 39, no. 5, Pp: 4760–4768, 2012.
[5]Bekkerman R and M. Gavish, "High-precision phrase-based document classification on a modern scale," in Proc. 11th ACMSIGKDD Knowl. Discovery Data Mining, Pp: 231–239, 2011.
[6] Cao. G J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in Proc. Annu. Int.ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Pp: 243–250, 2008.
[7] Chandrashekarand.G, F. Sahin, "A survey on feature selection methods," inComput. Electr. Eng., vol. 40, Pp: 16–28, 2014.
[8] Croft.B, D. Metzler, and T. Strohman,"Search Engines: Information Retrieval in Practice. Reading", MA, USA: Addison-Wesley, 2009.
[9] Debole.F and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets,"J. Amer. Soc. Inf. Sci. Technol., vol. 56, no. 6, Pp: 584–596, 2005.
[10]Eisenstein.J, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in Proc. Annu. Int. Conf. Mach. Learn, Pp: 274–281,2011.
[11] G.Forman, "An extensive empirical study of feature selection metrics for text classification," inJ. Mach. Learn. Res., vol. 3,Pp: 1289–1305, 2003.
[12] Gao.Y, Y. Xu, and Y. Li, "Topical pattern based document modeling and relevance ranking," in Proc. 15th Int. Conf. Web Inf. Syst. Eng., Pp: 186–201, 2014.
[13] Geng.X, T.-Y. Liu, T. Qin, A. Arnold, H. Li, and H.-Y. Shum, "Query dependent ranking using k-nearest neighbor," inProc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Pp: 115–122,2008.
[14] Heui Lim, "Improving kNN Based Text Classification with Well Estimated Parameters", LNCS, Vol. 3316, Pp: 516 – 523, Oct 2004.
[15] Huang. Y.H and S.-Y. Lin, "Mining sequential patterns using graph search techniques," inProc. Annu. Int. Conf. Comput. Softw.Appl., Pp: 4–9, 2003.
[16] Ifrim.G, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, Pp: 354–362, 2008.
[17] Kavitha Murugeshan, NeerajRK "Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm" IJITEE ISSN: 2278-3075, Volume-2, Issue-6, Pp: 28-29, May 2013.
[18]. L. P. Jing, H. K. Huang, and H. B. Shi. "Improved feature selection approach tf*idf in text mining." International Conference on Machine Learning and Cybernetics, 2002.
[19] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.
[20] Novovicova J., Malik A., and Pudil P., "Feature Selection Using Improved Mutual Information for Text Classification", SSPR&SPR 2004, LNCS 3138, Pp: 1010–1017, 2004.
[21]P. Jackson and I. Moulinier. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing Company, Amsterdam/Philadenphia, 2002.
[22]R. Sharma and S. Raman. "Phrase-based text Representation for managing the web document". In Proceedings of the International Conference on Information Technology: Computers and Communications (ITCC), Pp: 165–169.
[23] S.Charanyaa and K.Sangeetha, "Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 2, ISSN(Online): 2320-9801, February 2014.
[24]T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". In Machine Learning: ECML- 98, 10th European Conference on Machine Learning Pp: 137–142,1998.
[25]V.Vapnik, C. Cortes "Support-vector networks. Machine Learning", 20(3),Pp:273–297, September 1995.
[26]Y.Li and N.Zhong. "Mining ontology for automatically acquiring web user information needs". IEEE Transaction on Knowledge and Data Engineering, 18(4): Pp: 554-568.
[27]Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering". Proc. ACM 7th Conf. Information and Knowledge Management (CIKM '08), Pp. 1023-1032, 2008.
[28] 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.
[29] 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.

[30] Warnars.S "Mining Frequent Pattern with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP)",IEEE(ICoICT), Pp:87-95, 2014.

S.Brindha received B.Sc degree in Physics from Bharathiyar University. She done her Master Degree in Information Science and Management in Periyar University and she awarded M.Phil Computer Science from the Bharathiyar University. She has 3 years of teaching experience and 5 years of Technical Experience in Hash Prompt Softwares Pvt. Ltd. Currently She is doing her Ph.D computer Science in Erode Arts and Science College. Her Research area includes Data Mining and Text Mining.

Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 28 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 50 M.Phil research Scholars in various fields and guided 5 Ph.D Scholars. Currently he is Guiding 5 M.Phil Scholars and 8 Ph.D Scholars. He is a member of Board studies of various Autonomous Colleges and Universities. He published around 68 research papers in national and international journals and conferences. His current research interests include Image processing and Data Mining, Networking.