An Advanced Fuzzy Constructing Algorithm for Feature Discovery in Text Mining

Evana Ramalakshmi, Subhakar Golla
Dept. of Computer Science & Engineering.
Swarnandhra College of Engineering &Technology
JNTUK, East Godavari, India.

ABSTRACT
It is a big task to provide the accuracy of discovered relevance features in text documents for describing user requirements. Classification of data is biggest issue in more text documents because they have large number of words and data patterns. Most existing popular methods are used by word-based approaches. Still, they have all suffered from the problems of relevance and uncertainty. Over the years, there has been pattern-based methods should perform better result than word-based methods in describing user requirements. But, how to effectively use large scale patterns remains a typical problem in text mining. To overcome this problem, Fuzzy Relevance Feature Discovery Algorithm (FRFDA), classification techniques have been developed for relevance feature discovery. It describes both higher level and low level features based on word patterns. It is also classifies words into categories and updates those word weights based on their relevance and dispensation in patterns. The experimentation result proves that, the proposed FRFDA is better than existing manual and automation methods. The data set Reuters-21578 shows that the proposed model significantly outperforms faster and obtains better extracted features than other methods.

Keywords
Text mining, fuzzy similarity, feature clustering, text feature extraction, text classification, Fuzzy Relevance Feature Discovery (FRFD), Reuters Corpus Volume (RCV).

1. INTRODUCTION
The main theme of similarity feature discovery is to find out the useful information placed in large text documents, including both relevant and irrelevant features for describing text mining and classification results. Feature clustering is a biggest issue in naive information retrieval systems from theoretical and practical aspects also. Feature discovery text mining is also biggest problem in Web applications, and has received great attention from researchers in Text Mining, Artificial Intelligence, Information Retrieval and Pattern Recognition. Now a day's pattern mining is facing so many issues to find out relevance features in relevant and irrelevant documents [1]. Many patterns consists a good meaning for topic, but they have low support or frequency. Whenever minimum support is decreased then a lot of noisy patterns may be discovered. The second issue is the 'support' and 'confidence' problem, these are used in pattern mining but they did not supported to solving the problems in some of the cases. Generally frequent pattern mining is one of the good retrieving techniques for relevant and irrelevant documents. But the biggest problem is extraction of accurate weighted features. Such type of problems may solve by Pattern taxonomy mining (PTM) models have been proposed [2], [3]. It is a sequential pattern mining, identifying of white space is a term in sentences and paragraphs. Other hand, natural language processing (NLP) techniques are identified concepts in sentences. Previously many researchers proposed only term based techniques, but pattern based techniques are very beneficial to both relevant and irrelevant documents also [4]. Over the last decade, researchers have implemented several word-based techniques for assigning a score in documents because they used for filtering of information and text classification. Recently, many naive approaches were introduced for text categorization. Basically word based methods are used in relevant and unlabelled documents. At first step, it utilized a Rocchio classifier to retrieve a set of reliable irrelevant documents from the unlabeled set of documents [5]. At second step, a SVM classifier is used for classifying of text documents. Some of the unlabeled documents used both word based models as well as pattern based models as a rough analysis [6], [7].

In text categorization, the dimensionality of the feature vector is generally high. Reuters 21578 top-10, 20 Newsgroups and WebKB are three real-world data sets, they have more than 20,000 features. That is a main task for reducing of high dimensionality in to low dimensionality by using different classification algorithms. To overcome such type of problems, feature reduction techniques are introduced before text classification [8]. In general, both feature selection and feature extraction approaches are used for reduction of features. But feature extraction methods are high effective and computationally expensive than feature selection methods. So the main demand of large documents is efficient feature extraction algorithms. That's why we propose a fuzzy relevance feature discovery algorithm (FRFDA), which is an advanced feature clustering method to reduce the number of features for the text categorization. All terms in the feature vector of a document set are described as partitions, and processed one by one. Words that are similar to another cluster then those are placed in to same cluster, otherwise it is act like as an individual cluster. Every cluster is characterized by a membership function with statistical mean and deviation. After completion of all comparisons the number of clusters formed automatically, each cluster has one extracted feature. That extracted cluster refers a weighted combination of the words contained in the document [9].

The paper is organized as follows: In the next section, some related concepts about extraction of relevance features, In section 3 provides utilization of proposed algorithm. In section 4 experiment results of proposed algorithms FRFDA is presented. In section 5 describes conclusions.

2. RELATED WORKS
Different feature clustering methods have been introduced and studied for artificial intelligence applications. They may be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches [10], [11]. The

embedded techniques incorporate relevance feature selection as a part of the training process and those are usually specific to given learning algorithms, so they may be more accurate than the other three categories. Recent machine learning algorithms like decision trees or artificial neural networks are best examples of embedded techniques [12]. The wrapper approaches are using for the predictive accuracy of a predetermined learning algorithms, to determine the best of the feature selected subsets. The performance and accuracy of the learning algorithms is very high in large documents. Moreover, the generality of the selected features is low and the computational complexity is high. The filter approaches are not dependent on learning algorithms, as per good generality. But their computational complexity is less, but the accuracy of the learning algorithms is not guaranteed. Generally hybrid methods are treated as the combination of filter and wrapper approaches. Filter approach will reduce search space in large documents that will be considered by wrapper approaches [13],[14]. The combination of filter and wrapper approaches is to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter approaches. The wrapper approaches are computationally high and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Therefore, all comparisons methods shows filter approach is best one.

Relevance feature selection is a good method for selection of subset features from large datasets for modelling of machine learning systems. From last several years, a variety of feature selection methods (e.g., Filter, Wrapper, Embedded and Hybrid approaches, and unsupervised or semi-supervised methods) have been introduced in different fields. Feature selection is also one of the best methods for text mining and information filtering. It is the task of assigning documents to predefined classes. Several classifiers such as Naive Bayes, Rocchio, kNN, Fuzzy, SVM and Lasso regression have been implemented, in addition many believe that fuzzy is also a promising classifier. The main problem of classification is single class and multi-class problem. The best solution for multi-class problem is to divide it in to some independent binary classifiers then those will be classifies either relevant category or irrelevant category. Most of the feature selection methods are used the bag of words representation to select a set of features for the multi-class problem. So many feature selection methods for text classification, including document frequency (DF), term frequency (TF), inverse document frequency (IDF), information gain (IG), mutual information (MI), Compactness, First Appearance (FA) [15].

This paper describes, mainly focusing on relevant feature selection in large text documents. Relevant feature selection is a biggest issue on web search because it always compares with relevant document for user query. Still, the traditional feature selection methods are not suitable for selecting text features for solving relevance issue because relevance is a single class problem. The efficient way of feature selection for relevance is based on a feature weighting function. A feature weighting function indicates the rank of information indicated by the feature occurrences in a large document and reflects the relevance of the feature in that. Majority of word-based ranking models include tfidf based techniques [16], [17].

3. FUZZY RELEVANCE FEATURE DISCOVERY ALGORITHM

In this section, we introduce the FRFDA model for relevance feature discovery, which deals the relevant features in document with three groups: positive words, general words and negative words based on their compactness in a training set. Here the first discussion is the concept of 'specificity' in terms of the relative specificity" in training datasets and the exact 'specificity' in test datasets. We also present a way to understand whether the proposed relative 'specificity' is reasonable in term of the exact "specificity". After that it concludes word weighting method in this model.

3.1 The Categorical Approach

The partition (P1, P, P2) is used to clearly separates irrelevant documents from relevant ones. Consider two functions f1 and f2, f1 (w) is the approximate average weight of word for all relevant documents f2 (w) is the approximate average weight of word for all irrelevant documents. The partition (P1, P, P2) works as follows.

$$\int_{w1}^{wn} (f_1(w)-f_2(w))dt.$$

It is to make positive specific features separates far away from negative specific features. Suppose few words refers the same specificity then score of the cluster will be in crease. Where specificity function represented as a distance function. After completion of distance calculation then all words is separated by three categories. Therefore we required a clustering method to group all words in to three categories automatically by using the specificity function. So far, we assign some of the words that appear only in irrelevant documents in to the negative specific category P2, all remaining words Wi act like as own cluster Ci. The cluster Ci has two specifications those are maximum specificity and minimum specificity. Maximum specificity refers smallest specification value of elements in Ci and minimum specificity refers largest specification value of elements in Ci [18].

Let Ci and Cj be two clusters. The difference between the two clusters is defined as follows:

$$Dif\ (C_i, C_j) = Min\ \{|max_{spe}(C_i)-min_{spe}(C_j)|, |max_{spe}(C_j)-min_{spe}(C_i)|\}$$

If the difference is low between two clusters then we can combine them by using bottom-up approach. Suppose Ck be the combined cluster then it is described as Ck=Ci ∨ Cj. The combining method continues until three clusters are left whenever the number of initial clusters is more than three. The three clusters refers distances between two adjacent clusters in the retained three clusters should be greater than or equal to any other distances between two adjacent clusters. Which cluster has the highest minspe is determined as P1, the cluster has second highest minspe will form category P and the remains will be part of P2.

3.2 The Proposed Approach

There are some issues pertinent to most of the existing feature clustering methods. First, the parameter k, indicating the desired number of extracted features, has to be specified in advance. This gives a burden to the user, since trial-and-error has to be done until the appropriate number of extracted features is found.

The following algorithm describes for feature clustering in large datasets.

Feature Clustering Algorithm

Input: Features F, F1 and F2

Output: Three categories of words P, P1 and P2.

Method:

Step1. Initially all word categories are empty.

Step2. If any word Wi belongs to Feature F then it will place in to discovered feature F. Otherwise it will be compared to remaining test sets.

Step 3. Find maximum and minimum specificity of that word to compared sets.

Step 4. Separates positive, general and negative features.

Step 5. If it is belongs to any other three features category then send to those appropriate places.

Step 6. Calculate the percentages for feature clustering.

Step7. Repeat these steps to all sentences and topics until to reach the feature clustering.

Suppose, we are given a document set D of n documents d1, d2; . . . , dn, together with the feature vector W of m words w1, w2, . . . ,wm and p classes c1, c2, . . . ,cp, as specified. We construct one word pattern for each word in W. For word wi, its word pattern xi is defined as

$X=< X_{i1},X_{i2},...,x_{in}>$

$=<P(C_1|W_i), P(C_2|W_i),...., P(C_p|W_i)>,$

Where $P(C|W)= \sum^{n}_{q=1} d_{qi} * s_{qj} / \sum^{n}_{q=1} d_{qi}$

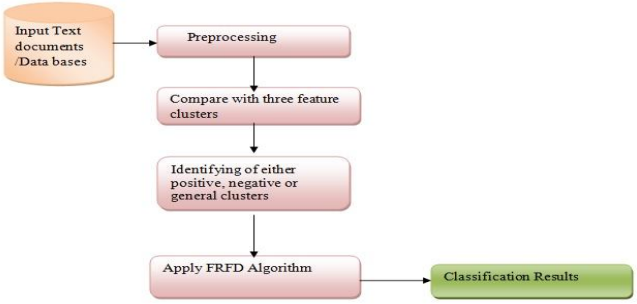

Fig.1. Architecture for Fuzzy Relevance Feature Discovery Algorithm

Second, when calculating similarities, the variance of the underlying cluster is not considered. The distribution of the data in a cluster is an important factor in the calculation of similarity. Third, all terms in a cluster have the same degree of contribution to the resulting extracted feature. Some cases, it may be better if more similar words are allowed to have bigger degrees of contribution. Our feature clustering algorithm is proposed to deal with these issues [19].

4. RESULTS AND DISCUSSION

The investigation of the proposed method is describing word classification could help to improve the best performance, this model is discussed in terms of the following approaches.

The FRFDA model classifies words into three clusters like general, positive negative specific terms by using the specificity function.

Case1. The specificity function firstly refers words into most documents.

Case2. The positive specific terms are referred as exact users wanted words, general terms are the necessary information for describing what users want. By using three categories together mostly we can get the best performance. Fuzzy relevance feature discovery (FRFD1) is a model for information filtering. It holds satisfactory performance for a given testing set [19].

Case3. FRFD2 overcomes the limitation of FRFD1 by using a clustering method to classify the words into three categories directly. It can achieve a almost similar performance as FRFD1. The RFD2 model also shows remarkable performance compared with the other models.

4.1 FRFD2 vs FRFD1

FRFD1 has two low level parameters like L1 and L2, those are using for low level words into three clusters. As per manual testing sets these parameters have low level values i.e., L1=0.1 and L2=0.2. FRFD2 model uses feature clustering for to automatically generation three categories P, P1and P2.

Figure.1 shows the average results of the five appearances on all 50 assessing paragraphs in large data sets, where %chg denotes the percentage change of FRFD2 over FRFD1. Here, FRFD2 can produce almost the same performance as FRFD1. In addition, a small improvement to four measures top _ 20, B=P, IAP and Fb¼1 except MAP was observed. These results fully supported to case3.

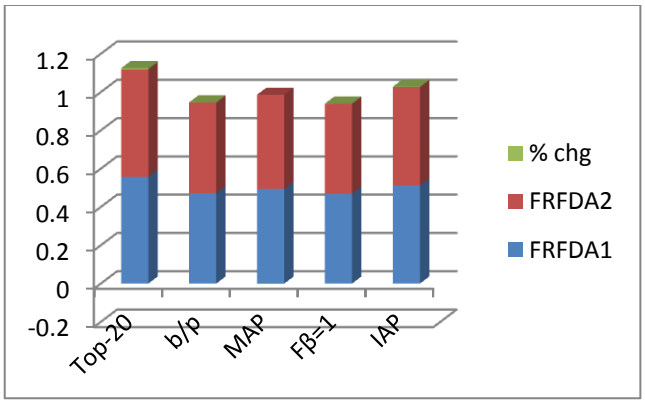

Fig.2. Comparison between FRFD1 and FRFD2 Models in all assessing paragraphs.

4.2 FRFD2 vs Pattern-Based Methods

This is the comparison between all pattern models with fuzzy relevance feature discovery model on RCV1 data base. We mentioned here some of patterns and n-grams like CBM etc [20]. The results on data collection RCV1 for all model in the first category (RFD2, language models (n-grams), CBM and other pattern-based models) are presented in Here change refers the percentage change of RFD2 over PTM. Basically, pattern-based methods struggle a lot in some topics as too much disturbance is occurred in the discovery of positive patterns in RCV1. Mostly closed sequential patterns perform better results than other patterns, and PTM deploying method outperforms greatly closed sequential patterns. In figure 2 FRFDA2 shows out standard performance results than PTM also.

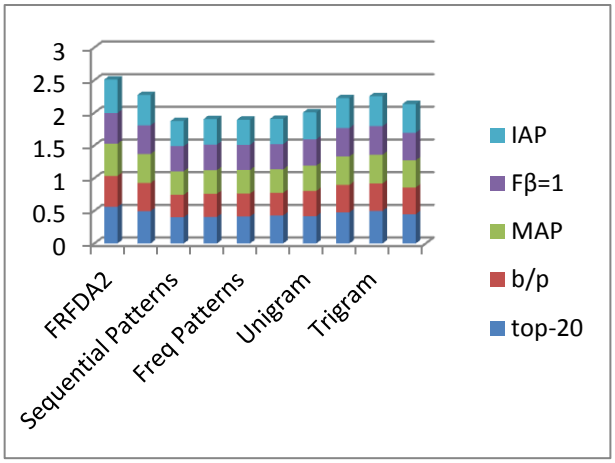

Fig.3 Comparison of all pattern based methods on RCV1

Figure 4 describes, simultaneously to see the effectiveness of using both positive and negative clusters for relevance feature discovery, we can compare FRFD2 with the best pattern based model PTM which uses positive patterns only in Reuter's data base.

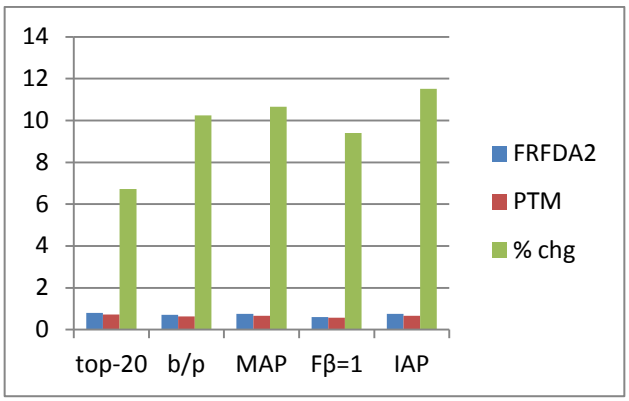

Fig.4 Comparison of FRFDA2 with PTM on Reuters database

4.3 FRFD2 vs Feature Selection Methods

The introduced method using FRFD2 already compared with main feature selection models including Rocchio, BM25, SVM, MI, x2 and Lasso. Figure 5 describes the experimental results on RCV1for all 50 assessing paragraphs are given.

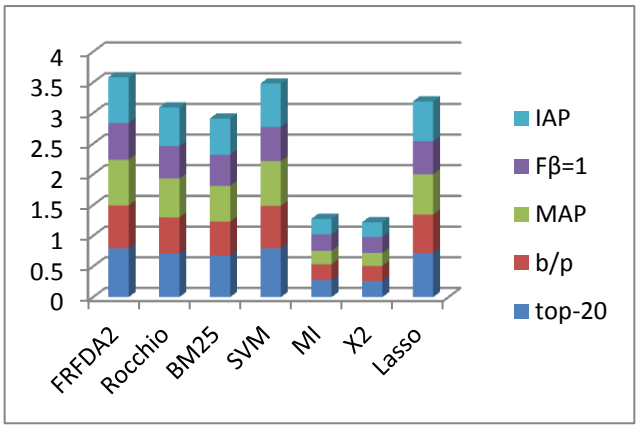

Fig.5 Comparison Results of All Models on RCV1

The proposed naive model FRFD2 achieved the best performance for the assessor topics, where FRFD2 is compared with Lasso (which is the second best term-based model on RCV1).

The average percentage of improvement over the standard measure is 8% percent with a maximum of 10.87% percent and minimum of 5.62%.

SVM is the second best word-based model on Reuters- 21578 data base. Figure 6 shows the FRFD2 is compared with SVM, it achieved the best performance than SVM. The maximum percentage of improvement over the Fβ=1 measure is 7.72%.

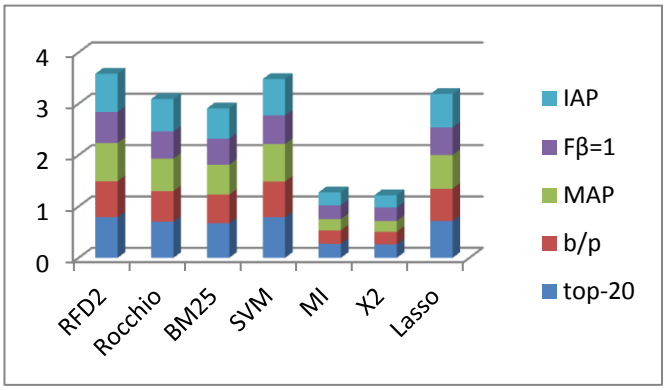

Fig.6 Comparison of all models on Reuter's data base.

5. CONCLUSION

This research proposes a fuzzy relevance feature discovery (FRFD) algorithm, which is a naive clustering approach to reduce the high dimensionality of the features in to text categorization. Some features relevance to each other is grouped into the same cluster. Such cluster is characterized by a membership function with statistical mean and deviation. If a term is not relevance to any existing cluster, a new cluster is created for this term. Feature similarity between a word and a cluster is defined by both the mean and the variance of the cluster. This paper describes an alternative method for relevance feature discovery in large text documents. It refers an approach to find and classify low-level features based on appearances and specification of the word. It also handles a method to retrieve irrelevant documents for weighted features. Experimentally, the development of fuzzy model is proved that proposed specificity function is flexible and word classification can be effectively approximated by a feature clustering method. This FRFDA model uses two exact parameters to fix the boundary between the categories. It reaches the most expected results, still it requires the manually testing of a large number of different values of parameters. The naive model uses a feature clustering technique to automatically group words into the three categories. When compared with the first model, the second model is much more efficient and achieved the good performance.

6. REFERENCES

[1] H. Li, T. Jiang, and K. Zang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," T. Sebastian, S. Lawrence, and S. Bernhard eds. Advances in Neural Information Processing System, pp. 97-104, Springer, 2004.

[2] Datasets for single-label text categorizatio. Http://web.ist.utl.pt/~acardoso/data sets/, 2010.

[3] D.D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," J. Machine Learning Research, vol. 5, pp. 361-397, http:// www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf, 2004.

[4] N. Slonim and N. Tishby, "The Power of Word Clusters for Text Classification," Proc. 23rd European Colloquium on Information Retrieval Research (ECIR), 2001.

[5] M.C. Dalmau and O.W.M. Flo´ rez, "Experimental Results of the Signal Processing Approach to Distributional Clustering of Terms on Reuters-21578 Collection," Proc. 29th European Conf. IR Research, pp. 678-681, 2007.

[6] X. Wang, H. Fang, and C. Zhai, "A study of methods for negative relevance feedback," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 219–226.

[7] Rao, Gudikandhula Narasimha, and P. Jagdeeswar Rao. "A Clustering Analysis for Heart Failure Alert System Using RFID and GPS." ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I. Springer International Publishing, 2014.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," in J. Mach. Learn. Res., vol. 3, no. 1, pp. 1157–1182, 2013.

[9] G. Narasimha Rao, R. Ramesh, D. Rajesh, D. Chandra sekhar."An Automated Advanced Clustering Algorithm For Text Classification". In International Journal of Computer Science and Technology, vol 3,issue 2-4, June, 2012, eISSN : 0976 - 8491,pISSN : 2229 – 4333.

[10] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau, "A two-stage text mining model for information filtering," in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 1023–1032.

[11] C. D. Manning and H. Sch€utze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[12] S. E. Robertson and I. Soboroff, "The TREC 2002 filtering track report," in Proc. 11th Text Retrieval Conf., 2002.

[13] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[14] S. Shehata, F. Karray, and M. Kamel, "A concept-based model for enhancing text categorization," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2007, pp. 629–637.

[15] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," in IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[17] I. Guyon, C. Aliferis, and A. Elisseeff, "Causal feature selection," in Computational Methods of Feature Selection Data Mining and Knowledge Discovery Series, Boca Raton, FL, USA: CRC, 2007 pp. 63–85.

[18] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Feature-based classification of time-series data," in Information Processing and Technology, Commack, NY, USA: Nova, 2001 pp. 49–61.

[19] C. A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints," in Proc. SIAM Int. Conf. Data Mining, 2004, pp. 11–22.

[20] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," ACM Trans. Database Syst., vol. 27, pp. 188–228, 2002.