IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 21-25
www.iosrjournals.org

Information Extraction from Text Document using Pattern Mining and Feature Extraction Method

Miss. K. S. Hantodkar [1], Dr. S.S.Sherekar [2], Dr.V.M.Thakare[3]
[1](kanak55hantodkar@gmail.com, *Dept. of Comp. Sci. & Engg*, SGBAU, Amravati, India)
[2](ss_sherekar@rediffmail.com, *Dept. of Comp. Sci. & Engg*, SGBAU, Amravati, India)
[3](vilthakare@yahoo.co.in,,*Dept. of Comp. Sci. & Engg*, SGBAU, Amravati, India)

Abstract: *Text mining is a method that discovers interesting information in text documents. It is difficult to find accurate feature in given text document for users to find what they want. Developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets. Pattern mining techniques are used for finding appropriate features in both relevant and irrelevant documents. Pattern mining has been extensively studied in data mining communities for many years. This paper, discusses briefly five methods, i.e Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, Relevant feature discovery model. This paper proposes a method improves the performance of the pattern mining, feature extraction method is used for information extraction from given text document. The proposed methodology for pattern mining from text documents using feature extraction method is the combination of Fuzzy feature clustering, Learning phrase pattern, Low rank shared concept and Relevance feature discovery. This will help user to get the required information from the given text document which the user has asked for.*

Keywords— *Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, Relevant feature discovery model*

I. INTRODUCTION

The text classification is dimensionality of feature vector which is usually huge. Therefore, developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets. Typically feature extraction method aims to convert the original high-dimensional data set to lower-dimensional by projecting process using algebraic transformations. Text mining discovers interesting info in text documents. It is difficult to find correct feature in text documents to help users to find what they need. A typical text classification system is combination of feature extractor and classifier. Classifications have limited performance by features, and others may get benefit by combinations of words. Phrase pattern is method that n-grams to allow gaps between words, this makes easy to get non-local behaviors. Many text mining applications have high-dimensional feature spacing; hence it is difficult to collect sufficient data from different domain. Both instance and feature-level approach reduce the distribution gap between training and testing set to propagate the label information and effective in various applications. Pattern mining techniques are used for finding appropriate features in both relevant and irrelevant documents. Pattern mining has been extensively studied in data mining communities for many years. The big obstacle of pattern mining for text mining is how to effectively use both relevant and irrelevant feedback.

This paper, studies five methods i.e Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, Relevant feature discovery model. To improve the performance of the pattern mining on text documents feature extraction method is used for information extraction from given text document.

II. Related work

The study on mining of text documents discusses the most relevant mining techniques developed in recent years. In fuzzy self-constructing feature clustering (FFC) algorithm, an incremental clustering approach to reduce the dimensionality of the features in text classification. Similarity-based clustering is one of the techniques developed in machine learning research. FFC algorithm reduces the dimensionality of the features in text classification. Words that are similar to each other grouped into same cluster. By this, the user needs not to specify the number of extracted features in advance. FFC runs faster than other clustering methods and provides comparably good or better extract features for classification. Future research in this is done by considering the clustering method for faster recognition of required words [1].

Effective pattern discovery technique is a pattern deploying and evolving process, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. An effective pattern discovery technique first calculates specificities of patterns and latter evaluate term weights according to the distribution of terms in discovered patterns other than the distribution in documents. The advantage of this is that it improves the accuracy of evaluating term weights because discovered patterns are more specific than whole documents. This has a disadvantage that it cannot perform on very big data of very large organization [2].

Learning phrase pattern algorithm determines when word classes are useful in locations of a phrase pattern, allowing for variable specificity depending on the amount of labeled data available improves performance when adding the phrase patterns to the existing n-gram features. Phrase patterns are useful features for text classification, also an efficient way of learning discriminative phrase pattern features. The algorithm determines when word classes are useful in specific locations of a phrase pattern. The results show the robustness of phrase patterns given erroneous ASR output. That is of practical value because little human effort is needed to use this type of word class [3].

Low rank shared concept (LRSC) space for adapting text mining model is a domain adaptation method that extracts the shared concept space between the source domain with sufficient labeled data and target domain with unlabeled data. This method share space by a linear transformation and find optimal solution considering combination of two criteria: the empirical loss on the source domain and the embedded distribution gap between the source domain and the target domain. Information extraction is one of method of LRSC to extract precise text fragments, which are basically chunks of consecutive tokens, each field of interest from a semistructured text document. For example, extracting job related information from recruitment websites. LRSC minimizes the domain gap and empirical loss on the labeled data simultaneously. Exploration of domain knowledge for extracting more concepts is a possible direction for future work [4].

Relevance feature discovery in text documents is a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity. The goal of relevance feature discovery in text documents is to find a set of useful features, including patterns, terms and their weights, in a training set, which consists of a set of relevant documents, and a set of irrelevant documents. This helps in significant improvement of effectiveness. The results show that the model achieves the best performance for comparing with term-based baseline models and pattern-based baseline models. The disadvantage is that the selected term must be less than 300 lines so that the algorithm will work properly [5].

This paper studies five methods i.e Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, Relevant feature discovery model and these are organizes as follows. **Section I** Introduction. **Section II** discusses Related work. **Section III** discusses existing methodologies. **Section IV** discusses Analysis and Discussion. **Section V** discusses proposed method and. **Section VI** discusses expected results. Finally **section VII** Conclude this paper.

III. EXISTING METHODOLOGIES

Many mining methods have been implemented over the last decades. There are different methodologies that implemented for mining text documents i.e Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method, Relevant feature discovery model.

Feature clustering algorithm: This deal with following issues.

1 Self-Constructing Clustering*:* This clustering algorithm is an incremental, self-constructing learning approach. Word patterns are considered one by one.

2 Feature Extractions: By applying this algorithm, word patterns have been grouped into clusters, and words in the feature vector W are also clustered accordingly. For one cluster have one extracted feature.

3 Text Classifications: The similarity threshold is applied to clustering algorithm. Assuming k clusters are obtained for the words in the feature vector W. Then finds the weighting matrix T and convert D to D0. Using D0 as training data, a classifier based on support vector machines (SVM) is built [1].

Effective pattern discovery: It is a technique that first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative training examples to find ambiguous patterns and try to reduce their influence for the low-

frequency problem. This technique can improve accuracy of evaluating term weights because discovered patterns are more specific than whole documents [2].

Frequent phrase pattern mining: It is an unsupervised method for feature learning, so it is difficult to optimize for discrimination. Mutual information is the reduction of uncertainty in a random variable after observing another random variable. This uses: i) the mutual information of the phrase pattern to determine if a phrase pattern is discriminative, and ii) the upper-bound for mutual information against the threshold to determine if any extensions of the phrase pattern may be discriminative[3].

LRSC domain: It adapted as follows:
1. A domain adaptation framework for text mining problems. This discovers low-rank shared concept space where the empirical loss on the labeled data and the distribution gap between source and target domain are jointly minimized.
2. It can kernelize method in the RKHS so as to generalize the model by making use of the powerful kernel functions. The alternate optimization strategy can solve this model efficiently.
3. Theoretically analyze the expected error evaluated by common loss functions in the target domain under the empirical risk minimization framework, showing that the error bound can be controlled by the expected loss in the source domain and the embedded distribution gap.
4. Domain adaptation method is capable of considering multiple classes and their interactions simultaneously [4].

The RFD model: describes the relevant features in relation to three groups: positive specific terms, general terms and negative specific terms based on appearances in a training set. The goal of relevance feature discovery in text documents is to find a set of useful features, including patterns, terms and their weights, in a training set consists of a set of relevant documents, and a set of irrelevant documents [5].

IV. ANALYSIS AND DISCUSSION

Experimental results to show the effectiveness of fuzzy self-constructing feature clustering method. Three data sets for text classification research: 20 Newsgroups, RCV1 and Cade12 were used in experiments for feature clustering. Then compared with other three feature methods: IG is one of state of art feature selection approaches, IOC is an incremental feature extraction algorithm, and DC is a feature clustering approach. FFC can run much faster than DC and IOC in feature reduction also provide comparably good or better extracted features for classification [1].The effectiveness of PTM (IPE) to find the correlation between achieved improvements and the parameter, giving the ratio of number of negative documents greater than threshold to the number of all documents. As a result, PTM (IPE) is the method that uses the least amount of patterns for concept learning compared with others. This is due to the efficient scheme of pattern pruning is applied to the PTM (IPE) method. The inner pattern is deploying strategy that provide an effective evaluation for reducing the side effects of noisy patterns [2]. Phrase patterns improve the performance using reference transcripts. Phrase patterns with word classes usually work better than without word classes. The experiments show the performance of phrase patterns that can be improved with properly chosen word classes [3]. LRSC adopt the recall, precision, and F-measure as the evaluation metrics. Recall is the number of articles that correctly classified divided by the actual number of articles in each class. Precision defined number of articles that correctly classified divided by number of all the articles predicted same class. F-measure defines harmonic mean of recall and precision. LRSC method minimizes the domain gap and empirical loss on data simultaneously [4]. In RFD offender selection play important role for giving negative feedback in process of feature discovery and deploying. It is believed that negative feedback has some useful information that helps to identify the boundary between relevant and irrelevant information for improving the effectiveness of relevance feature discovery. It is able to balance the percentages of positive specific and general terms for large reduced noises [5].

Text mining Techniques	**Advantages**	**Disadvantages**
Fuzzy self-constructing Feature Clustering	FFC method runs much faster than other classification method in feature reduction.	When a document set is transformed to a collection of word patterns relevance among word patterns can be measured.
Effective Pattern Discovery Technique	It improves the accuracy of evaluating term weights.	It cannot perform on very big data of very large organization.
Learning	Efficient solution is obtained by this	Context-dependent word classes

Discriminative Phrase Pattern	algorithm with selection criterion.	perform does not perform well.
Low-Rank Shared Concept	Method is capable of conducting the domain adaptation task.	Exploration of domain knowledge for extracting more discriminative concepts is not done.
Relevant feature discovery	Effective use of both relevant and irrelevant feedback to find useful features is done; and pattern features together rather than using them in two separated stages.	The selected term must be less than 300 lines so that the algorithm will work properly.

TABLE 6: Comparisons between Fuzzy self-constructing Feature Clustering, Effective Pattern Discovery, Learning Discriminative Phrase Pattern, low-rank shared concept, Relevant feature discovery method.

V. PROPOSED METHODOLOGY

Many mining strategies have been used, such as the Fuzzy self-constructing Feature Clustering method, Effective Pattern Discovery Technique, Learning Discriminative Phrase Pattern method, low-rank shared concept method and Relevant feature discovery model, each of which has its own special characteristics. In text clustering, selecting important feature is important, which has critical effect on output of clustering algorithm. Fuzzy feature clustering algorithm for text classification that reduces the dimensionality of features in text classification. Due to this, words that are similar to each other are grouped into the same cluster. Learning phrase pattern algorithm determines word classes that are useful in specifying locations of a phrase pattern. Phrase patterns are useful features for text classification and it is an efficient way of learning discriminative phrase pattern features. Low rank shared concept (LRSC) space for adapting text mining model is a domain adaptation method that extracts the shared concept space. Information extraction is one of method of LRSC to extract required text fragments for each field of interest from a semi structured text document. Relevance feature discovery in text documents is a method to find and classify low-level features based on both their appearances in the higher-level patterns and their specificity. Relevance feature discovery in text documents finds a set of useful features, including patterns, terms and their weights, in a training set, having set of relevant documents and irrelevant documents. Hence, the proposed methodology for pattern mining from text documents using feature extraction method is the combination of Fuzzy feature clustering, Learning phrase pattern, Low rank shared concept and Relevance feature discovery. When a text document is given, user will be asked for information which he required. Then the given phrase will be checked in the text document b using Learning phrase pattern algorithm. This gives the specific location of the given phrase. Then by using Low rank shared concept, information is extracted from the sentence here the phrase is present. Relevance feature discovery algorithm determines whether the given information is relevant text or not. Finally the information got from the given text is arranged according to their weights. Hence the required information is extracted from the given text document.

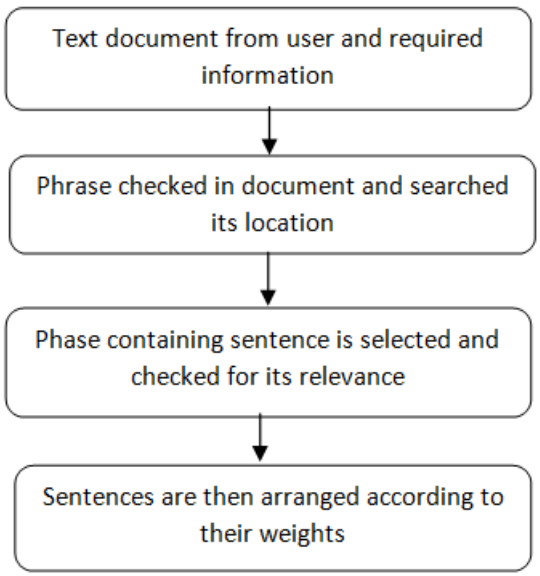

Fig. Proposed Framework

VI. Expected RESULTS

The expected result for this proposed method will be that, it will give efficient information from the given text document given by user. This method will first ask user for his required information. Then the given phrase is checked in the text document. After locating the phrase the sentence is selected and is checked whether it is relevant information or not. Then the sentences are arranged according to their weights. This proposed method is more efficient than other information extraction methods applied earlier. This method consumes less time than any other method.

VII. CONCLUSION

This paper focused on efficient method for information extraction from given text document. This paper proposed a method, which is combination of Learning Discriminative Phrase Pattern method, low-rank shared concept method and Relevant feature discovery model that will help user to extract information from given text document.

The method proposed for information extraction will derive better and efficient result in terms of retrieval time and the contextual text which will be more appropriate as compared to the existing methodologies.

VIII. FUTURE SCOPE

From Observation, the scope to be studied in future work, the propose method can be added more efficient clustering methods that will support various text documents and also will reduced computation time.

REFERENCES

[1]. Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee," A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", *IEEE Transaction on Knowledge And Data Engineering, VOL. 23, NO. 3, MARCH 2011.*

[2]. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,"Effective Pattern Discovery for Text Mining," *IEEE Transaction on Knowledge And Data Engineering, VOL. 24, NO. 1, JANUARY 2012*.

[3]. Bin Zhang, Alex Marin,Brian Hutchinson and Mari Ostendorf, "Learning Phrase Patterns for Text Classification", *IEEE Transactions on Audio, Speech, AND Language Processing, VOL. 21, NO. 6, JUNE 2013.*

[4]. Bo Chen, Wai Lam, Ivor W. Tsang, and Tak-Lam Wong, "Discovering Low-Rank Shared Concept Space for Adapting Text Mining Models," *IEEE Transactions on Pattern Analysis AND Machine Intelligence, VOL. 35, NO. 6, JUNE 2013.*

[5]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, "Relevance Feature Discovery for Text Mining," *IEEE Transaction on Knowledge And Data Engineering, VOL. 27, NO. 6, JUNE 2015*