BAG OF WORDS INVOLVING SELECTION OF TERM SETS USING APRIORI ALGORITHM IN TEXT CLASSIFICATION

R. Iswarya#1 V.Pandiyarajau#2 A.Kannan#3

PG Student#1 PhD Scholar#2 Professor and Head#3

Department of Information Science and Technology
College of Engineering, Anna University
Chennai, India.

Abstract- **Text mining area enables people to extract the relevant information from the given text. This approach is advantageous, because it allows the users to retrieve the useful information and helps to avoid unambiguity in the text. Text Mining tasks are classified as text classification, text clustering and summarization of documents. Phrases are important in the field of text mining and information retrieval. Phrases identification, classification of phrases constitute of a major importance because, phrases are constructed by means of term sets. Term sets or item sets constitute a set of two terms usually bigram. The occurrence of a term in the document can be found by means of providing binary weights. Document representation can be done with Bag of Words (BOW) model. The main motivation of this paper is that term sets are constructed by computing the frequency of each term in the corresponding document and weighting for a term is also provided. Terms sets involving adjacent pair and non-adjacent pair are taken into consideration and this is used for the classifying positive documents and negative documents. Association rule mining is used for the construction of term sets. News Group dataset was taken which consists of twenty thousand messages is one which is widely used. In this paper, first pre-processing was done by means of stop word removal and stemming. Finally, association rules are formed by using apriori algorithm and term sets are formed.**

Keywords: Text Mining, Bag of Words, information retrieval, association rules, apriori

I. INTRODUCTION

Text mining makes use of the automated methods for exploiting the enormous amount of knowledge available in text documents. It is considered as an emerging and exciting area in the field of computer science which helps to establish a strong relation between texts. The purpose of text mining is to analyze the text where a form of intelligence is applied which can be used in the information retrieval systems. When the user gives a query and performs a search in search engine, the relevant information has to be retrieved to the user. Hence, text classification is considered as an important parameter and widely used in the retrieval of text. Text classification involves classifying a text into pre-defined category. Since, the texts involved are of unordered collection, the relevant text has to be mined and extracted. Since the phrases are discriminant in nature, generalizations of n-grams constitute a phrase.

Text Mining represents significant step in the retrieval of text. Text mining tools could be technologies which are capable of answering sophisticated questions and performing text searches with an element of intelligence. Typical text mining tasks include text categorization, text clustering, and production of analyzing the sentiments, summarizing the documents, and entity relation modelling. To make text mining tasks successful, the presented texts should be analyzed using natural language processing (NLP) techniques. The uses of natural language processing techniques enable text mining tools to get closer to the semantics of a text source. One word may refer to different many meanings and one phrase could be interpreted in many ways. This is important, especially when the text mining tool is expected to discover knowledge from texts. Thus natural language processing is a foundation for text mining, and it becomes a critical part of a text mining system. Precise information can be extracted using the concept of words, phrases and keywords. However, the main challenging issue in NLP is the natural language is always ambiguous. The advantage of this text mining is that the relevant data gets extracted drastically thereby making search engines retrieve the highly meaningful information.

II. RELATED WORK

The literature study in this section provides details about the n-gram based approach, term set construction, selecting the relevant feature and rule based extraction and bag of word model techniques are discussed.

2.1 N-gram based approach

Phrases and their construction is considered as a crucial part in text mining. N-gram approach was proposed by Furnkranz [3] for classifying the phrases. N-gram means, it is a combination of one or more words. It includes monogram, bigram and trigram. The advantage of using N-gram approach is many term sets can be constructed. Phrase patterns are sequence of words which also involve sequence of terms which are adjacent to each other. Identifying the phrases and their usage are discriminant in nature was proposed by Bekkerman et al. [2]. The phrases construction and their uses according to its context varies accordingly was proposed and the term set is used for selecting the terms which play an important role in the field of information retrieval.

2.2 Term set Construction

The selection of term set using apriori algorithm was proposed by Dimad Badawi et al. [1]. Term sets can be effectively constructed using association rule mining. Bigrams can be constructed easily, but when the size of ther term gets increased by trigram, four gram, five gram, the size of the complexity gets increased. Text Classification task gets improved when phrase patterns are used. The advantage of using term set based approach is that, the query gets retrieved according to the content in information retrieval system. In this work, in addition to the phrases, the use of term sets also plays a major role. Phrase Construction and their usage based on the content seem to be the most challenging part as far as NLP and Information Retrieval are concerned. The concept of phrases and their identification is dealt here. The disadvantage of using statistical based approach is that the phrases are not classified correctly. Automatic text classification helps to classify the text and this is used to support systematic reviews. In this work, instead of incorporating unigrams with bigrams, the bigrams alone were used separately. It is shown that unigram doesn't produce adequate features than bigrams. So, the enhancement of text and their features lies in the hands of bigrams.

2.3 Selecting relevant features

SVM is one of the classification algorithm used for text categorization. Joachims [4] proposed support vector machine for selecting the relevant features for choosing the term sets. The important property of SVM is that it selects the relevant features by discarding the irrelevant feature. It is observed that SVM, a machine learning algorithm works fine for text classification task when compared to naive bayes classifier.

2.4 Rule based extraction

Text mining technique using association rules to extract prepositions. The use of prepositions in a sentence should also be projected with less ambiguity. The phrase in the context not only contains words or terms, but also includes prepositions. Parts of Speech tagging plays a crucial role in the field of text mining to classify the terms according to the context was proposed by Toutanova et al. [5]. The Parts of Speech tagging tags each sentence according to the context specified. The use of incorporating prepositions by applying rule deduction is one of the ways of extracting phrases. Bag of Words (BOW) model is used for representing the documents. Machine learning techniques play an important role in classifying the text. One of the machine learning techniques includes support vector machine, so the use support vector machine in text mining was proposed by Sebastiani [6]. This destruction is caused because the words are not arranged properly. Moreover the phrase itself projects the information about the topic. Weighting factor is provided in this approach. The representation of documents as vectors possesses Term Frequencies (TF).The word TF- stands for term frequency, which means the number of times a word occur in a document and Inverse Document Frequency(IDF) helps to find the term which is most frequently occurring or less frequently occurring. Syntactic approach is used to identify the phrases based on a grammatical relation. Here, the type of phrases includes Noun Phrase, Verb Phrase, and Adjective Phrase. Classifications of features which involve selection of terms play an important role in the field of information retrieval. So the classification of features for the term set was proposed by Scott.S et al. [7] .Term sets can be constructed by using apriori which involves two steps. Based on threshold value, the support and confidence values are calculated. Syntactic phrases helps the user to classify the text based on grammatical relation, it does not yield a better scoring when selecting a subset of phrases and also it becomes a failure when rare distinct phrases are used. The result showed in this paper proves that when syntactic phrases are used with BOW means, the accuracy would be improved.

2.5 Bag of word Model

The phrases based methods are discussed above. Cohen et al. [11] proposed the bag of word model and also says the phrase which deals with semantic relations which involve synonyms and hyponyms. Bag of word method aim at finding the single term for a particular document. The frequent less words are removed from the corpus and terms which possess greater frequency are taken into consideration for training the corpus. The observation made in this BOW model is that user should correctly understand the text. Words, Sentences and their relationship in the text should reach the human easily. So, this type of phrases and their phrases based representation i.e.) BOW model is widely useful for interpreting the text.

The discussions on noun and noun phrases were proposed by lewis et al. [12]. Noun Phrases are of wide importance in some context. This phrase is usually followed by noun or adjectives. A key phrase is different from noun phrases. This keyword phrases is used in extracting the relevant meaning according to the context. The extracted keyword phrase should be highly relevant to the user query and highly meaningful. This key phrase should reveal the exact gist of the content about a particular topic. The concept of word net was proposed by Miller et al. [13] .This tool acts as a thesaurus for finding the synonyms and also tries in fetching the exact relationship between words. This word net helps in preventing the sense of disambiguity between words and text according to a context. The advantage of word net is that the synonymy of a word can be found and related accordingly. If the phrases change, correspondingly their meaning change which in turn destroys the meaning of context.

From the above proposed techniques, the bigrams are constructed using apriori algorithm [10] and it is advantageous because the term sets help in conveying the terms that belong to positive documents and also negative documents.

III. SYSTEM ARCHITECTURE

The architecture of the system proposed in this work consists of three major components namely pre-processing, construction of term sets using apriori, extraction of features, selecting the features and classifying the algorithm based on support vector machine as shown in the Figure 1.

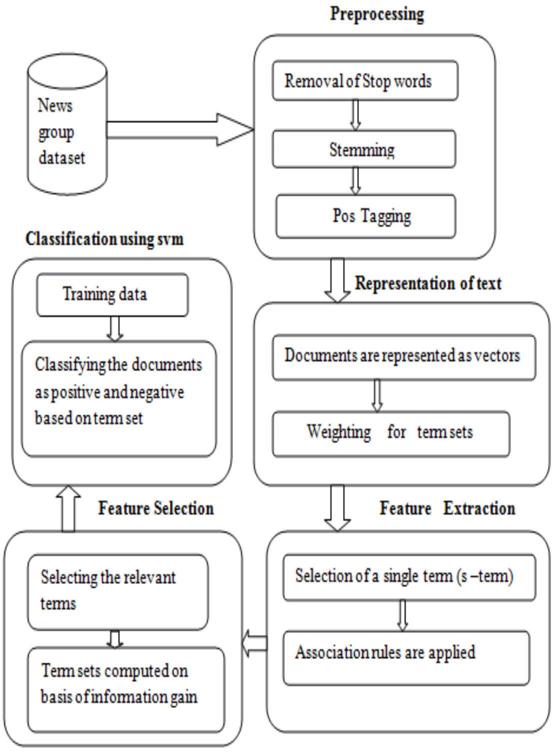
Preprocessing
News group dataset
Removal of Stop words
Stemming
Pos Tagging
Classification using svm
Training data
Classifying the documents as positive and negative based on term set
Representation of text
Documents are represented as vectors
Weighting for term sets
Feature Selection
Feature Extraction
Selecting the relevant terms
Term sets computed on basis of information gain
Selection of a single term (s –term)
Association rules are applied

Figure. 1 Frame work for Bag Of Word Model

The functional modules are defined as follows, namely pre-processing of electronic reviews, representation of text, extraction of features based on term sets, selection of features and classification using support vector machine. The input for pre-processing step is set of electronic reviews. The steps are illustrated as follows,

1) Pre-processing step involves stop word removal, stemming and pos tagging. The most commonly used words in English are worthless. These are termed as stop words. E.g.: this, a, an, the.

2) Stemming is used to stem or find out the root word. Porter Stemmer algorithm [9] is used for stemming. E.g.: walkingness, walking, walked is stemmed to a root word “walk”.

3) Parts of speech tagging (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word.

4) Document representation is done using Bag of Word Model by computing frequency of each term. This is termed as text representation for the documents.

5) Association rule mining is constructed using apriori algorithm was proposed by Omar et al [10]. These rules are formed based on threshold value.

6) Support vector machine is the classification algorithm and it is computes the correctly classified terms according to its content.

Feature Selection involves selection of suitable term. This feature selection technique involves mutual information and information gain. These techniques are useful in determining imbalanced text classification. So the selection of term set to extract the relevant feature was proposed by Ogura et.al [8].The suitable terms means selecting a term in such a way that the relevance factor for that term is high.SVM classification algorithm is used to classify the correctly retrieved terms and also used to compute the less number of times a term is used.

IV. PROPOSED SYSTEM

Apriori algorithm is used for construction of term sets was proposed by Osmar et al.[10].In this work, this algorithm is used for feature selection and the concept used is association rule mining has been used. It is known that in rule mining, if the term is said to occur, then the corresponding subset of terms are also said to occur.. The apriori algorithm follows two properties viz, join property and pruning property. Join property is used to combine two terms i.e.) a term set of size 2. Pruning property helps to remove the irrelevant subsets based on minimum support.

Applying this algorithm, the association rules are formed. The rules generated will depend on the number of terms which has high frequency. With the help of weka tool, association rules can be formed. The rules generated will be ten and based on support value corresponding term sets will be formed.

The input to the apriori will be set of documents and term sets will be the output.

Algorithm: Association rules using apriori

```
1:  for each set of terms in Document $D_j$ ,
2:  calculate frequency of a term $F_i$
3:  check if the term frequency is greater than threshold
4:  calculate the frequency of terms for next set of documents,
5:  Ignore the terms which has less threshold value
```

```
6:  candidate item sets are formed for frequency of terms
7:  end for
8:  finally term sets are constructed
```

Step 1: Frequent Item Sets

The frequent item sets are found from a set of documents. The frequent item sets determine how frequently a term is occurred in each and every document. The threshold value will be fixed initially.

Step 2: Join Step and Prune Step

If a term computed is found to be less than the minimum support value, that terms will be pruned in the first transaction itself. It will be taken to further evaluation. Other terms which has value greater than the minimum support will be moved to next transaction.

Step 3: Candidate Item Sets

The candidate item sets are formed by combining the frequent item sets which are common to both the documents. Likewise, for frequent 1 item set, candidate 1 item set will be generated and for frequent 2- item set , candidate 2 item set will be generated and so on.

Step 4: Generation of Association Rules

Association rules are formed finally and confidence values for each term are computed effectively. Rules generated will be more if and only if the frequency of the term sets in the document is high. Otherwise, rules generated will be very less. If there is a less occurrence of terms, then the confidence value is also found to be very less.

Table 1 Weighting for the terms in documents

Terms	Doc1	Doc2	Doc3
resistor	1	0	0
circuit	1	1	0
Device	1	1	1

As the table 1 indicates, the terms used are resistor, circuit and device which belongs to electronic reviews which is taken from newsgroup dataset. The weighting (1, 0) indicates which means 1 indicating the term is present and has occurred in the document and 0 indicates the term does not belong to that document.

Dataset Used

This dataset is a collection of 20,000 messages, collected from 20 different net news groups. The 20 newsgroup collection has become a popular dataset for experiment in text mining applications of machine learning groups such as text classification and text clustering.

V. PERFORMANCE ANALYSIS

In this section, we discussed the performance analysis of our proposed system with the use of unigrams in the existing system. The existing system is implemented with set of unigrams and its efficiency is less and our proposed system is implemented with the construction of term sets (bigrams). Precision is the number of relevant documents retrieved by taking in to account the total number of retrieved documents which are found to be irrelevant. Recall is the relevant documents retrieved while classifying the algorithm. F-score is the weighted average of precision and recall. This recall value gets increased with respect to number of documents. The frequencies of terms are computed in each document and it is found the recall value gets increased by `increasing the number of terms in each document and also the corresponding term sets gets increased.

1) Terms and term sets:

If the number of term increases, the corresponding term set also gets increased. Suppose, if we consider 20 set of documents, for each and every document compute the frequency for each term in every document. If the occurrence of a term is found maximum in a document, then the weight for the term increases with respect to the document.

Figure 2 illustrates terms and their term sets. If we keep on increasing the terms with respect to the threshold, term sets also gets increased.

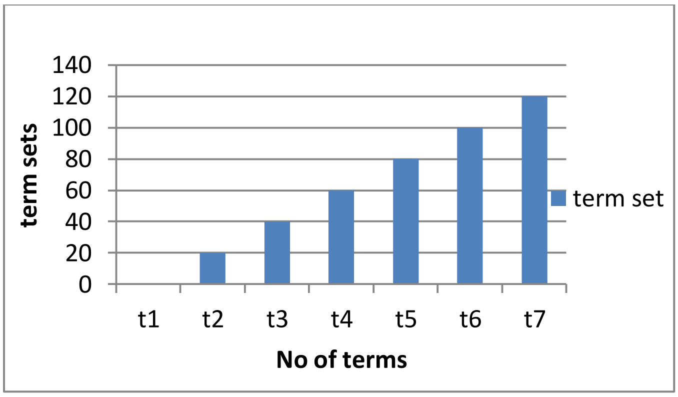

Figure 2. Construction of terms and term sets

2) Recall

Recall value gets computed with respect to number of retrieved documents which are relevant. If the number of document increases, correspondingly recall value gets increased and then gets decreased at a certain point. This is because the term frequency for a particular document may be high. The computation gets high if the document size increases. Figure 3.illustrates the recall graph is computed with respected to number of documents.

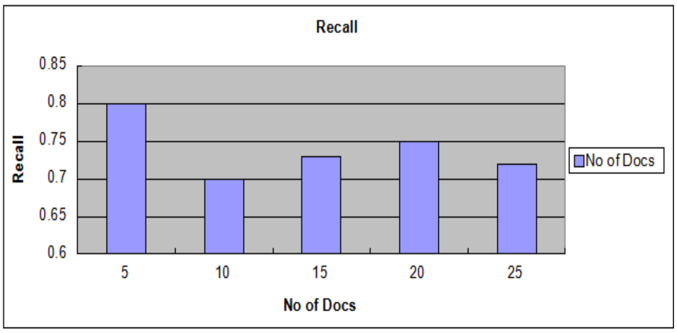

Figure3. Recall graph

VI. CONCLUSION

The term sets are constructed effectively using apriori algorithm. Extraction of term sets (bigrams) produced better result when compared to unigrams. The efficiency of bigrams gets improved by means of association rules. The two most important parameters support and threshold values helps in generating association rules which in turn is useful in construction of term sets.

VII. FUTURE WORK

In the future work, our goal is to extend the terms sets which are widely to be used in information retrieval system. When the user gives a query the corresponding term sets should be searched in such a way that it is relevant to the given context. The query retrieved during search should be relevant to the user. So, in order to avoid disambiguity between the terms, the effective term sets are constructed.

REFERENCES

[1] Dimad Badawi, Hakan Attncay, “A novel framework for termset selection and weighting in binary text classification”,“ Journal Of Engineering applications of Artificial Intelligence”,vol.35,no.2,pp.38-53,2014.

[2] Bekkerman.,Allan,“Using Bigrams in Text Categorization”, A Technical Report On Information Retrieval -408,” International Center of Intelligent Information Retrieval, UMass Amherst, vol.3,no.4,2004

[3] Furnkranz, J ,“A Study Using n Gram Features for Text Categorisation”, a Technical Report on Enterprise Application Integration,”Australian Research Institute for Artificial Intelligence, vol.2,no.3,1998

[4] Joachims, T. “Text categorization with support vector machines learning with many relevant features”.In Proceedings of the 10th European Conference on Machine Learning, Springer-Verlag, vol.3,no.2, pp 137–142, 1998.

[5] K.Toutanova, D.klein, C.D.Manning, and Y.Singer, “Feature –Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” in the Procceedings of North American Chaper of the Association For Computation Linguistics , vol .3, no.2, pp 12-16, 2009.

[6] Sebastiani, “Machine learning in automated text categorization” Association for Computing Machinery” In the Proceedings of Advanced Computing Machinery,vol.34 ,no.4,pp 1-47,2002.

[7] Scott, S.,Matwin, S., “Feature engineering for text classification”. In the Proceedings of the 16th International Conference on Machine Learning (ICML-99), vol.3,no.4, pp. 379-388, 1999.

[8] Ogura,H, Amano,H., Kondo, M ,Comparison of metrics for feature selection in imbalanced text Classification, Expert System Application, vol.38,no.5,pp 78- 89, 2011.

[9] Porter M.F, “An Algorithm for Suffix Stripping”, Program, vol.3,no.4, pp.130-137, 1980.

[10] Osmar, Antonnie, “Classifying text documents by associating terms with text categories,” Proceedings of the 13th Australian Database Conference”, Australian Computer Society, vol.5, pp 215-222, 2002.

[11] Cohen, William W. and Singer, Yoram , “ Context-Sensitive Learning Methods for Text Categorization. Special Interest group on Information retrieval , vol.*96*. pp. 307-316,1996.

[12] Lewis, David D. and Sparck Jones, Karen, “Natural Language Processing for Information Retrieval. Communications of the Association Computation Linguistics, vol.39., noo.2, pp.. 92-101.,1996.

[13] Miller, George A.. “Word Net: an On-line Lexical Database. International Journal of Lexicography, vol.3no.4,pp235-244,1990.