IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. V (Jan. 2014), PP 34-38
www.iosrjournals.org

Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature

Soumya George K[1], Shibily Joseph[2]
[1]*(Department of Computer Science and Engineering, Government Engineering College, Thrissur, India)*
[2]*(Department of Computer Science and Engineering, Government Engineering College, Thrissur, India)*

Abstract : *Text classification is the task of assigning predefined categories to free-text documents based on their content. Traditional approaches used unigram based models for text classification. Unigram based models such as Bag Of Words(BOW) models are not considering co-occurrence of set of words in a document level. This paper proposes a way to find co-occurrence feature from anchor text of wikipedia pages, proposes a way to incorporate co-occurrence feature to BOW model. Finally the method is analyzed to know how it performs in task of text classification.*

Keywords: *Text Classification, Natural Language Processing, Machine Learning, Bag Of Words, Naive Bayes classifier*

I. Introduction

Obtaining information resources relevant to an information need from huge amount of information available in digital form is called Information Retrieval. Efficiency of an IR system is about how intelligently it provides information to users. For example, searches are based on indexing. Good indexes are provided by Text classification(TC).

Text classification is the task of assigning predefined categories to free-text documents based on their content. The classification task is also called text categorization. Many classification tasks have traditionally been solved manually. Books in a library are assigned Library of Congress categories by a librarian. Book was not available in electronic form. So a computer was not needed for classification. Until the late 80's the most popular approach to TC was a knowledge engineering (KE). In KE, a domain expert defines a set of rules to classify documents under the given categories. Current approach to solve TC is Machine Learning(ML). Machine Learning methods are used for automated TC.

Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data ie, a domain expert is replaced by computer. It can say a partial yes answer to the question whether machines do what we can do. While building a ML system, one have to take design decisions on algorithmic approach, data representation, computational efficiency, and quality of the resulting program.

Machine learning algorithms can be organized into following taxonomy based on the way of learning.
1) Supervised learning generates a function that maps inputs to desired outputs from labeled examples.
2) Unsupervised learning models a set of inputs, like clustering from unlabeled examples.
3) Semi-supervised learning combines both labeled and unlabeled examples to generate an appropriate function or classifier.

A typical TC process consists of the following steps: preprocessing, dimensionality reduction, and classification. Different approaches have been experimented for all these steps [2]. In preprocessing, text is converted to a form suitable for training. Normally stopwords such as pronouns, prepositions, conjunctions are removed. One can use the Porter stemming algorithm in this step [3]. During training, input set of data is used to discover potential features of categories. Finally a test-set is used to test the strength of features discovered during training. A number of supervised TC techniques have been explored in the literature including centroid-based approaches [7], regression models [8], SVM [9] nearest neighbor classifiers [10], decision trees [4], Bayesian classifiers [11]. Unsupervised techniques include Clustering [5] and Hidden Markov Models [11]. This paper mainly discusses on supervised TC.

The Bag Of Words(BOW) model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text is represented as an unordered collection of its words, disregarding grammar and even word order. In case of text classification, a word in a document is assigned a weight according to its frequency in the document and frequency in between different documents. Words together with their weights form BOW. This work introduces a new feature to BOW model. Also evaluates the proposed scheme to know how it performs in task of text classification.

1.1 Example showing BOW model

The following models a text document using BOW.

The”:1	”Sun”:2	”is”:3	”a”:4	”star”:5	”beautiful”:6	”Moon”:7	”satellite”:8
1	2	2	1	1	1	0	0

1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |

Here are two simple text documents D1 and D2:

1 2 3 4 5 6

D1: “The Sun is a star. Sun is beautiful.”
D2: “The Moon is a satellite.” 8
7
Based on these two text documents, a dictionary is constructed as:
{”The”:1 ”Sun”:2 ”is”:3 ”a”:4 ”star”:5 ”beautiful”:6 ”Moon”:7 ”satellite”:8 } → Dictionary with index
Documents have 8 distinct words. Each document is represented as an 8-element vector

D1 D2

[1, 2, 2, 1, 1, 1, 0, 0] [1, 0, 1, 1, 0, 0, 1, 1] → freq for D1 & D2

where each entry of the vectors refers to count of the corresponding entry in the dictionary.

II. Related Works

Traditional text classification methods considered only terms ie, unigrams features [1]. Text was represented as Bag Of Words(BOW) using unigrams. But in unigram based representation, there will not be any consideration for relation between unigrams.

One solution to the above problem is usage of word pairs [12]. Koster and Seutter used a feature induction method which involve combination of single words and word pairs. In the method, nouns are extracted with their modifiers. Phrases are represented by an abstraction called Head/Modifier pairs. Rather than just throwing phrases and keywords together, they start with pure HM pairs and gradually add more keywords to the document representation. They use the classification on keywords as the baseline, which is compared with the contribution of the pure HM pairs to classification accuracy, and the incremental contributions from heads and modifiers. The authors show that when using pairs without BOW the results of classifiers decreases. And when used both pairs and BOW the results improved.

Usage of WordNet is a solution to BOW approach, since it considers relations like hypernym, synonym and antonym [13]. WordNet groups English words into different sets. WordNet groups English words into different sets. It describes experiment using a new representation of text based on WordNet hypernyms.
The algorithm for computing hypernym density requires three passes through the corpus.
1) During the first pass, assigns a part of speech tag to each word in the corpus.
2) During the second pass, all nouns and verbs are looked up in WordNet and a global list of all synonym and hypernym synsets is assembled. Infrequently occurring synsets are discarded, and those that remain form the feature set. A synset is defined as infrequent if its frequency of occurrence over the entire corpus is less than 0.05N, where N is the number of documents in the corpus.
3) During the third pass, the density of each synset is computed for each example resulting in a set of numerical feature vectors. Density is defined as the number of occurrences of a synset in the WordNet output divided by the number of words in the document. It lead to more accurate and more comprehensible rules. But it is not considering co-occurrence of words. Few efforts have been taken to incorporate background knowledge into document representation.

Maciej Janik and Krys Kochut presented a method using ontology, which is using semantic graphs for classification task [14]. It discusses on converting a text document into a thematic graph of entities occurring in the document, ontological classification of the entities in the graph, and then determining the overall categorization of the thematic graph, and as a result, the document itself. In the presented method, the ontology becomes the classifier. The approach tries to find a semantic similarity between a document and some fragment of the ontology that describes a certain category. To achieve it, the analyzed document must be transformed into a structure that is similar to the ontology. This process focuses on the creation of a semantic graph from the document, and employs entity matching and relationship identification. The semantic graph is then used to measure the document’s semantic similarity to the categories defined in the ontology. It does not require a document training set. In a context of supervised machine learning, this methods is not applicable.

III. Motivation

In bag of words (BOW) model, words in a document are used as basics for representing that document. Word Co-occurrence, a form of word order is not considered in the model. By co-occurrence, we mean frequent occurrence of terms from a text corpus alongside each other. Obviously the feature is possible only for a set of words, whose length is greater than or equal to two. Also occurrence of features in a document level is considered ie, whether terms occur together in a document or not.

Consider the bigram ”computer science”. When one see the word ”computer”, there is a chance that one will see the word ”science” next to it. As explained before, WordNet is considering only predefined set of relations. None of relations using WordNet can represent the bigram ”computer science”. Co-occurrence of words is not incorporated with BOW approach. A bigram of words is a co-occurrence of 2 terms, where one term occurs next to the other and is only a subset of co-occurrence.

Suppose we have a set of three terms. Possible combinations that can be formed from them with length greater than or equal to two is 4. It gives 4 possible co-occurrence. When number of terms increases, analysis becomes tedious. Considering whole vocabulary of corpus is difficult. So one obvious question is how can we find some selected sets for finding combinations. It will be easy, if we can find some good set of words in the beginning. Then one can form combination of terms from the selected set and co-occurrence of each combination is checked in the corpus. This paper checks whether wikipedia pages can be the source of finding co-occurring terms. Also investigates the usefulness of anchor texts in this context. Wikipedia is a collaboratively edited, free Internet encyclopedia supported by the non-profit Wikimedia Foundation. Wikipedia's 26 million articles in 286 languages are written collaboratively by volunteers around the world.

This work
1) propose a way to find co-occurrence feature from anchor text of wikipedia pages
2) propose a way to incorporate co-occurrence feature to BOW model.
3) analyzes the proposed scheme to know how it performs in task of text classification.

IV. PROPOSED SCHEME

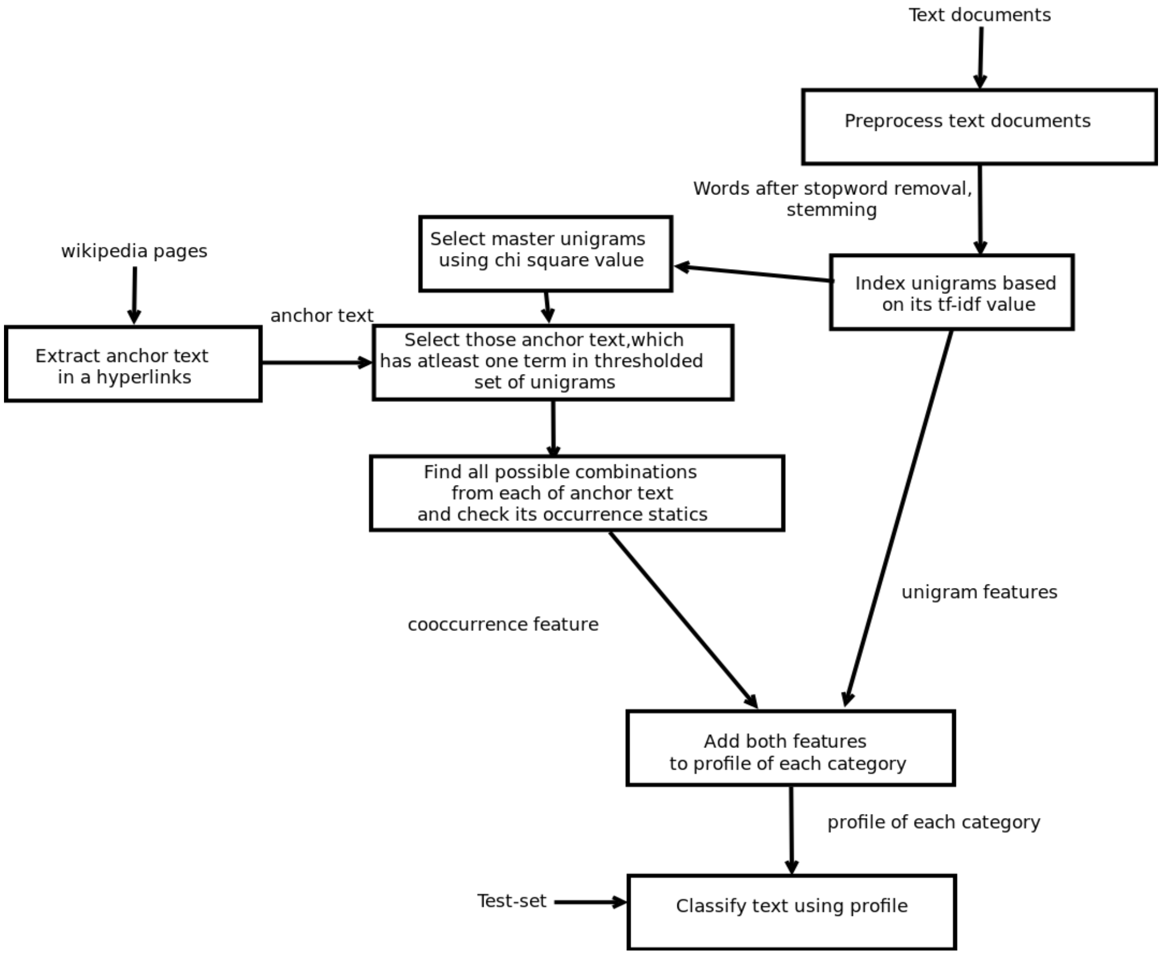
Text documents
Preprocess text documents
Words after stopword removal, stemming
Select master unigrams using chi square value
wikipedia pages
Index unigrams based on its tf-idf value
anchor text
Extract anchor text in a hyperlinks
Select those anchor text,which has atleast one term in thresholded set of unigrams
Find all possible combinations from each of anchor text and check its occurrence statics
unigram features
cooccurrence feature
Add both features to profile of each category
profile of each category
Test-set
Classify text using profile

Figure 1: Proposed scheme

Fig.1 gives an overview of the steps involved. One has to extract features from two different kind of inputs: raw text and sample wikipedia pages. Steps include the following: Finding unigram features, Indexing of unigram features, Feature selection on unigram features, Filtering of anchor text, Indexing of co-occurrence feature, Text classification using a supervised learning algorithm.

4.1 Finding unigram features

First, unigram tokens are extracted from text. Then unigrams stemmed afterwards using the Snowball stemmer. Stopword are removed. Techniques like stemming help in compensating for data sparseness. Words are tokenized using regular expression.

4.2 Indexing of unigram features

The two main components that affect the importance of a term in a document are the Term Frequency (TF) factor, Inverse Document Frequency (IDF) factor. Term frequency of each word in a document (TF) is a weight which depends on the distribution of each word in documents. It expresses the importance of the word in the document. Inverse document frequency of each word in the document database (IDF) is a weight which depends on the distribution of each word in the document database. It expresses the importance of each word in the document database . A weight is assigned to each unigrams using its TF-IDF.

4.3 Feature selection on unigram features to find master unigrams

The chi-square statistic is a different way to compute the lack of independence between the word and a particular class. This score can be used to select those features with the highest values for the chi-square statistic. This function weeds out the features that are the most likely to be independent of class and therefore irrelevant for classification [6]. Unigrams selected from this step is called master unigrams. Master unigrams are only used for filtering anchor text ie, whole vocabulary of corpus will be contributing to form Bag Of Features(Bag Of Features is explained in 4.6).

4.4 Filtering of anchor text

Next we will consider the input of wikipedia pages. Extract Anchor text from wikipedia pages. Then select those anchor text, which has at least one term in important set of unigrams got from previous step. This will be an effective method of filtering. Obviously we have to consider co-occurrence of those terms, which have a high chance of occurrence. It has the chance of greater utility than other terms. The unigram we used for selection of anchor text is called a ”master unigram”. Filtering of anchor text help for noise removal.

4.5 Indexing of co-occurrence features

This model is on the assumption that all combinations of words formed from each of the anchor texts, whose length is greater than two has a high chance of co-occurrence in a document level. Combinations from anchor texts form the candidate space for co-occurrence feature. Those candidates which prove occurrence in the training data will get used to augment the BOW model. Document frequency thresholding is used as a feature selector for co-occurrence. Document frequency of co-occurrence feature is the number of documents in which a feature appears. Remove those features whose document frequency is less than a predetermined threshold. The set of words corresponding to a co-occurrence feature is found. To index co-occurrence features extracted, frequency of co-occurrence feature and inverse document frequency of co-occurrence feature will get used.

4.6 Text classification using a Naive Bayes classifier

Using two different kinds of features ie, unigram features and co-occurrence features a Bag Of Features is formed. The naive Bayes classifier combines with a decision rule. The most probable class is assigned as the class of test document. We have C is a class variable and F is the feature variable. The corresponding classifier, a Bayes classifier, is defined as follows:

$$classify(f1,\dots,fn)=argmaxP(C=c)\prod_{i=1}^{n} p(Fi=fi/C=c)$$

V. EVALUATION

In this section, we evaluate our augmented BOW model on the TC task by comparing performance with BOW model. A balanced corpus 20-newsgroup dataset is used. The performance metrics used in the experiments is F1 score. F1 is a combined form for precision(p) and recall(r), which is defined as the harmonic mean of precision and recall. Equation is given by:

$$F1=2\cdot\left(\frac{p\cdot r}{p+r}\right)$$

We used F1 to evaluate the classification performance for individual category.

5.1 Datasets and Experimental Settings

The 20 Newsgroups data set divided almost evenly among 20 different UseNet discussion groups. 20Newsgroups dataset is downloaded from http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz. It has a standard train/test split. The task is to classify an article into the one newsgroup (of twenty) to which it was posted. Stemming is done using snowball stemmer. Features are indexed. ”alt.atheism”, ”comp.graphics”,”rec.sport.hockey”, ”sci.electronics” are used for evaluation. These classes are selected because they have related wikipedia pages.

5.2 Effect on F1-score

To give a more visualized comparison, Fig.2 give the distributions of F1 scores for the 20-newsgroup corpus. We can see that for all classes the F1-score improved.

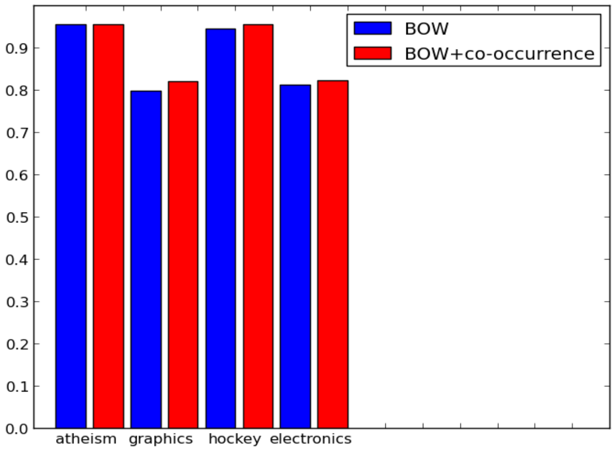

Figure 2: Comparison of F1 score

VI. Conclusion

This paper considers the usefulness of wikipedia anchor text in augmenting BOW model. It proposes an idea how this anchor text can be utilized to form a co-occurrence feature in context of text classification. The result shows that co-occurrence feature in document level has helped classification and is able to show an improvement in text classification. Co-occurrence may also provide good indexes in Information Retrieval.

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful suggestions that helped in improving the quality of this paper.

REFERENCES

Journal Papers:

[1] Fabrizio Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (2002), no. 1, 1–47.

[2] Kjersti Aas and Line Eikvil, Text categorisation: A survey, Raport NR 941 (1999).

[3] Martin F Porter, An algorithm for suffix stripping, Program: electronic library and information systems 14 (1980), no. 3, 130–137.

Books:

[4] G Luger and WA Stubblefield, Artificial intelligence: Strategies and structures *for complex problem solving, 1998.*

[5] John A Hartigan, Clustering algorithms, John Wiley & Sons, Inc., 1975.

[6] Richard Lowry, Concepts and applications of inferential statistics, R. Lowry, 1998.

Proceedings Papers:

[7] A class-feature-centroid classifier for text categorization, Proceedings of the 18th international conference on World wide web, ACM, 2009, pp. 201–210.

[8] David D Lewis and William A Gale, A sequential algorithm for training text classifiers, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., 1994, pp. 3–12.

[9] Thorsten Joachims, A statistical learning learning model of text classification for support vector machines, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 128–136.

[10] Piotr Indyk and Rajeev Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM, 1998, pp. 604–613.

[11] Paolo Frasconi, Giovanni Soda, and Alessandro Vullo, Text categorization for multi-page documents: A hybrid naive bayes hmm approach, Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, ACM, 2001, pp. 11–20.

[12] Cornelis HA Koster and Mark Seutter, Taming wild phrases, Advances in Information Retrieval, Springer, 2003, pp. 161–176.

[13] Sam Scott and Stan Matwin, Text classification using wordnet hypernyms, Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference, 1998, pp. 38–44.

[14] Maciej Janik and Krys J Kochut, Wikipedia in action: Ontological knowledge in text categorization, Semantic Computing, 2008 IEEE International Confer- ence on, IEEE, 2008, pp. 268–275.