Volume 3, Issue 1, January 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Incorporating Distributional Features on phrases in Text Categorization

Sravan Yadav Eadala*
Assistant Professor (OG)
Dept. of Computer Science and Engineering
SRM University
India

Dr. M Janaki Meena
Professor
Dept. of Computer Science and Engineering
Vellammal College of Engineering
India

Abstract-- Text mining is the process of discovering new, previously unknown information, from a usually large amount of different unstructured textual resources. Text Categorization is the task of assigning predefined categories to natural language text. This process of Text Categorization comes in preprocessing stage of Text Mining process. Feature can be a unit or weight assigned to represent a document. Feature Selection is a technique of selecting subset of features that best derives to characterize a document. Features for Text Categorization could be done with words, phrases or sentences that occur in training documents. Using bag of words, abundant information cannot be represented fully, since features selected may be redundant and irrelevant. By considering statistical methods, better features could be selected, that are dependent to a category. Moreover, position of the appearances of features plays a vital role in selecting good features. So, the distributional features, which include compactness of the appearances and position of the first appearance, had been incorporated on statistical methods. In this paper, performance had been evaluated by incorporating distributional features on statistical methods and compared with other feature selection techniques, for both words as well as phrases.

Keywords-- Distributional features, Text Categorization, Data Mining, Text Mining and Statistical Methods

I. INTRODUCTION

Text Categorization has been attracting more attention from researchers due to its wide range of applicability. Many classifiers in Machine Learning have been applied to this task, like Naïve Bayes, Neural Network, Support Vector Machine and k-Nearest Neighbor [Bing and Zhou, 2009]. Feature Selection is the most important technique used in data preprocessing for mining the information. It reduces the number of features, removes irrelevant or redundant data and improves mining performance. Feature Selection is the process of selecting subset features from the original features. Feature Selection can be found in many areas of data mining like classification, clustering, association rules, etc [Liu and Lei, 2005]. Features for selection used are words as well as phrases. Using phrases seems to be an interesting idea, as they have smaller degree of ambiguity and good degree of robustness than words. A unigram is a stemmed word. An n-gram is an alphabetical ordered sequence of n unigrams obtained after performing stop word removal and stemming [Matwin and Sebastiani, 2001].

A number of approaches have been introduced in order to select good features for document categorization. Of all, a novel approach and all of the above mentioned classifiers were based on same text representation, "Bag of words," where a document is represented as a set of words appearing in that document [Bing and Zhou, 2009]. Values assigned to each word usually expressed how frequently the word appears. However, these values are not enough for the Text Categorization and the problem with this approach is that, it considers all features present in the document. Features, thus formed, may be redundant or irrelevant, some may misguide if there are more irrelevant features than relevant ones. In such a case, selecting a subset of original features often leads to better performance.

It is reported that feature selection can improve the efficiency and accuracy of text classification by removing redundant and irrelevant terms from the corpus [Luo and Chung, 2008]. Traditional feature selection methods for classification are either supervised or unsupervised, depending on whether the class label information is required for each document. Those unsupervised feature selection methods, such as the ones using document frequency and term strength, can be easily applied classifying. But it is shown that supervised feature selection methods using the $\chi 2$ statistic can improve the classification performance better than unsupervised methods when the class labels of documents are available for the feature selection [Luo and Chung, 2008]. In many previous text mining researches, the $\chi 2$ term-category independence test has been used for feature selection. By considering their $\chi 2$ statistic values, features that have strong dependency on the categories can be selected, which is called CHI method. This $\chi 2$ term-category independence test has been extended by introducing another statistical method that measures whether dependency between a term-category is positive or negative, which is called CHIR method [Luo and Chung, 2008].

Based on the word's distribution in the document, the distributional features can be measured based on its compactness of the appearances and position of first appearance of the terms [Bing and Zhou, 2009]. The compactness of the appearances measures whether the appearances of a word concentrate in a specific part of a document or spread over the whole document. In the former case, the word is considered as compact, while in the latter situation, the word is considered as less compact.

The second consideration is the position of the first appearance of a word. This consideration is based on an intuition that the author naturally mentions the important contents in the earlier parts of a document. Therefore, if a word first appears in the earlier parts of a document, this word is more likely to be important.

The rest of the paper is organized as follows. In section 2, all the feature selection techniques used have described. In section 3, the modified system design has explained. In section 4, with the experimental results, the performance of words as well as phrases is compared among the feature selection methods, also, a comparison on the precision and recall values and F-measure values have been tabulated. Section 5 contains conclusion of the work.

II. FEATURE SELECTION TECHNIQUES

A. χ2 TERM-CATEGORY INDEPENDENCE TEST

The degree of dependency between a term-category is by using $\chi 2$ statistic. This can be done by comparing the observed frequencies in 2-way contingency table with the expected frequencies. An example to explain $\chi 2$ independence test:

TABLE 1
A 2x2 Term-Category Contingency Table

	c	**~c**	∑
w	40	80	120
~w	60	320	380
∑	100	400	500

In order to analyze relation between a term w and category c, a 2-way contingency table should be created as showed above. The row variable, term, has two possible values, {w, ~w}. The column variable, category, may take either one in {c, ~c}. Each cell at the position (i, j), where i ∈ {w, ~w} and j ∈ {c, ~c}, contains the observed frequency, denoted by O (i, j). For instance, O (w, c) is number of documents which are in category c, contains term w and (~w, ~c) is number of documents which neither belong to c nor contain w. The $\chi 2$ term-category independence test is calculated as showed below. The expected frequency E (i, j) can be calculated as:

$$E(i,j) = \frac{\sum_{a\in\{w,\sim w\}} O(a,j) \sum_{b\in\{c,\sim c\}} O(i,b)}{n} \qquad (1)$$

The $\chi 2$ statistic is defined as:

$$\chi 2_{w,c} = \sum_{i\in\{w,\sim w\} j\in\{c,\sim c\}} (O(i,j) - E(i,j))^2 / E(i,j) \qquad (2)$$

For the example considered, the expected frequencies values are as follow:
E (w, c) = 24, E (w, ~c) = 96, E (~w, c) = 76, E (~w, ~c) = 304 and $\chi 2_{w,c} = 17.61$. The degree of freedom is (2 – 1) x (2 – 1) = 1. From the $\chi 2$ distribution table, the critical value is 10.83 and the features below this value are rejected and would be removed from the feature space [Luo and Chung, 2008].

B. TERM-CATEGORY DEPENDENCY MEASURE $R_{w,c}$

The above mentioned $\chi 2$ statistic test cannot express the document information fully, as it does not consider relevancy between term and category. The dependency between term and category may be positive or negative. This can be explained with the criterion called relevancy. This is denoted by $R_{w,c}$ and defined as follows:

$$R_{w,c} = \frac{O(w,c)}{E(w,c)} \qquad (3)$$

If this ratio is close to 1, then there will be no dependency between the term w and category c. If observed frequency is larger than expected frequency, then the ratio is larger than 1 and there will be positive dependency. If the ratio is less than one, then the dependency will be negative. So, if $R_{w,c}$ is larger than 1, the term w is relevant to category c and term is statistically significant. For the contingency table, Table 1, $R_{w,c} = 1.67$, thus the term is strong positive dependency on category c, even its $\chi 2_{w,c} = 17.61$.

C. DISTRIBUTIONAL FEATURES

The word's distribution in a document is modeled as number of appearances of the word in the corresponding part, which can be modeled as an array. The length of the array is the total number of parts. In information retrieval, there are three types of passages [Bing and Zhou, 2009]. Discourse passage is based on logic components of documents like sentences and paragraphs. Because of inconsistency in passage length this is avoided. Semantic passage is portioned according to contents. Since its performance is influenced by the effect of partition algorithm, it is also avoided. The third type is window based which is simply a sequence of words. By considering certain standard length of window, this type is simple to implement. Here, the part is based on window passage. The distributional features include compactness of the appearances of the word and the position of first appearance of the word below. The position of first appearance of a word t and compactness of appearances of the word t are defined as follows. For document d, sentences n and the array of words array (t, d) = $[c_0,c_1,\ldots..,c_{n-1}]$:

$$\text{FirstApp}(t, d) = \min_{i \in \{0\ldots n-1\}} c_i > 0 ? i : n \qquad (4)$$

$Compact_{PosVar}$: This defines the variance of the positions of all appearances. For this, mean position of all appearances is first calculated and then, the mean distance between the position of each appearance and mean position is calculated as position variance.

$$\text{Compact}_{\text{PosVar}}(t, d) = \frac{\sum_{i=0}^{n-1} ci * |i - \text{centroid}(t,d)|}{\text{count}(t,d)} \qquad (5)$$

$$\text{centroid}(t, d) = \frac{\sum_{i=0}^{n-1} Ci * i}{\text{count}(t,d)} \;\&\; \text{count}(t, d) = \sum_{i=0}^{n-1} Ci$$

III. SYSTEM DESIGN

The modified system that is designed is diagrammatically represented as showed below:

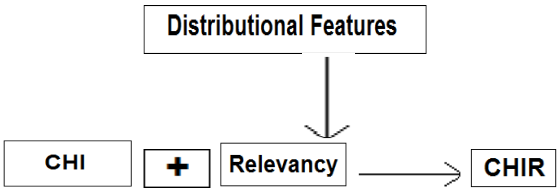

Fig. 1 Modified system abstract view

The system that developed is shown in an algorithmic form below:
Step1: Initially, the raw text documents are considered

Step2: As a step of preprocessing, the strings, words and phrases are tokenized.

Step3: The terms, thus formed are stemmed and in case of the phrases, sorting is done, so that, syntactically same phrases are indexed with the same prefixes.

Step4: For each term obtained after stemming and sorting, a contingency table is formed, based on term-category relation as showed in table 1.

Step5: Then, for each term, values are assigned by using $\chi 2$ statistic function that constructed based on observed and expected frequencies, showed in equation (2).

Step6: By using relevancy, from equation (3), the terms that are strong positive dependent towards the category are selected.

Step7: The terms that are not selected after step6, based on distributional features, from equation (4) and (5), values are assigned and terms beyond the threshold limit are neglected.

Step8: As a final step, the performance of the words and phrases, thus, obtained are calculated and that is compared to the performance of other feature selection techniques.

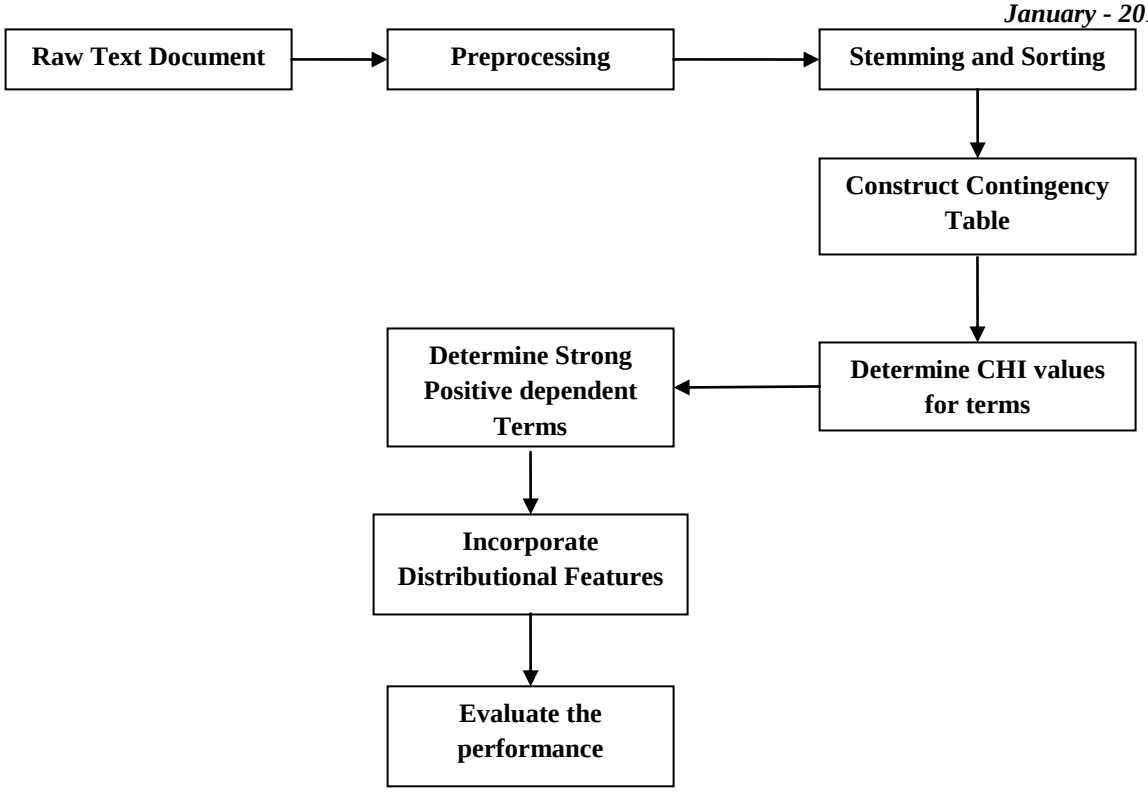

Fig. 2

Modified system Flow chart

IV. EXPERIMENTAL RESULTS

The experiments had been conducted in order to evaluate performance of words as well as phrases using Naïve Bayes classifier. The data set used for this is from 20Newsgroup, from which two thousand nine hundred and forty documents and six categories were considered. The total distinct words obtained are 27,123 and phrases are 60,697. Various threshold limits used are as follows: for $\chi 2$, the terms with the value greater than 10.83 were rejected and with value less than or equal to 10.83 were selected; for relevancy, if it is greater than one, terms were considered and less than or equal to one, were rejected; for the position of the first appearance, the terms that appear first 5% of the lines in the document were considered; for compactness, if the value is less than or equal to 11, terms were selected and beyond that, were rejected. The performance of words and phrases had been compared by using various feature selection techniques, using precision, recall and F-measure values, as showed below:

TABLE 2
Results based on precision and recall for words

	Bag of words		**CHI statistic**		**CHIR**		**Distributional Features**		**CHIR + Distributional Features**	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
alt.atheism	0.72	0.531	0.743	0.543	0.79	0.686	0.626	0.71	0.835	0.671
comp.windows.x	0.839	0.361	0.868	0.655	0.568	0.908	0.41	0.91	0.575	0.914
sci.crypt	0.843	0.678	0.791	0.533	0.834	0.626	0.848	0.41	0.853	0.639
sci.med	0.789	0.549	0.872	0.61	0.725	0.792	0.834	0.504	0.72	0.818
sci.space	0.753	0.449	0.795	0.545	0.81	0.72	0.756	0.639	0.798	0.708
soc.religion.christian	0.31	0.847	0.334	0.82	0.865	0.696	0.804	0.647	0.875	0.728

TABLE 3
Results based on precision and recall for phrases

	Bag of words		CHI statistic		CHIR		Distributional Features		CHIR + Distributional Features	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
alt.atheism	0.558	0.483	0.576	0.485	0.592	0.491	0.685	0.764	0.9	0.912
comp.windows.x	0.85	0.171	0.867	0.17	0.85	0.25	0.897	0.613	0.962	0.517
sci.crypt	0.965	0.557	0.96	0.48	0.962	0.68	0.645	0.915	0.94	0.945
sci.med	0.958	0.23	0.952	0.18	0.961	0.247	0.833	0.652	0.969	0.621
sci.space	0.92	0.462	0.928	0.48	0.937	0.461	0.883	0.837	0.948	0.92
soc.religion.christian	0.236	0.86	0.254	0.856	0.281	0.867	0.74	0.81	0.52	0.963

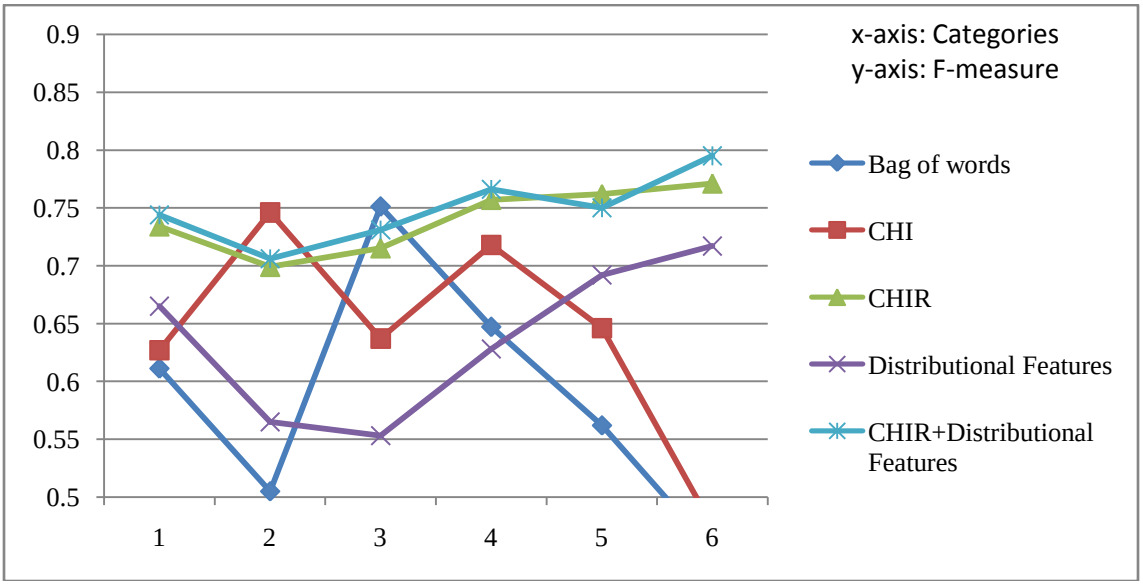

Fig. 3 Results based on F-measure for words

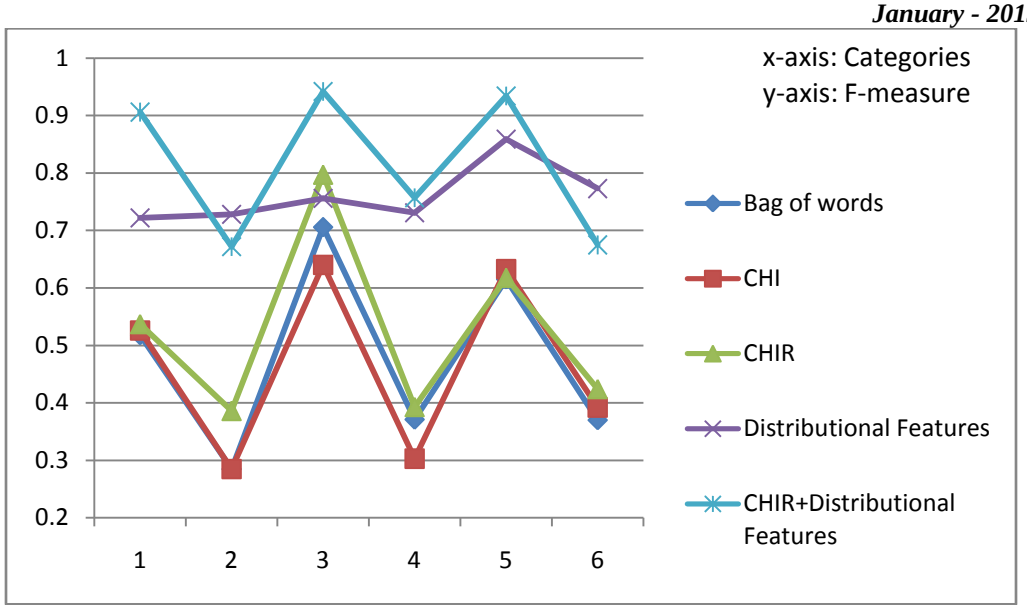

Fig. 4 Results based on F-measure for phrases

Thus, the results indicate that using CHIR along with distributional features yields better performance than other methods.

V. CONCLUSION

With the "bag-of-words" representation, previous researches usually assign a word with values that express whether this word appears in the document concerned or how frequently this word appears. Although these values are useful for text categorization, they have not fully expressed the abundant information contained in the document. With phrases, a structured concept can be expressed, they will have smaller degree of ambiguity, natural language processing technology allows good degree of robustness and a document can rank higher than a document with words. Thus, using supervised feature selection methods like statistical methods, the classification performance can be improved better than unsupervised methods, like bag-of-words.

Since, $\chi 2$ statistic method cannot address about the terms that are strong positive dependent, relevancy had considered. Also, position of the appearance of the terms plays a vital role for selecting good features for classification, distributional features had incorporated. The results reveal that, after incorporating distributional features on statistical methods for phrases, good results were obtained compared to the other features selection techniques. Moreover, the statistical methods are not affected by the size of the document, whereas, distributional features are more obvious for the larger documents.

ACKNOWLEDGEMENT

The authors want to thank all the staff and faculty members for their support and suggestions, Department of Computer Science and Engineering, PSG College of Technology as well as Department of Computer Science and Engineering, SRM University.

REFERENCES

[1] Xiao-Bing and Zhi-Hua Zhou, "Distributional Features for Text Categorization", IEEE Trans. Vol. 21, No: 3 pp:428 - 442, 2009.

[2] Yanjun Li, Congnan Luo, and Soon M. Chung, "Text Clustering with Feature Selection by using Statistical Data", IEEE Trans., 2008.

[3] M.F. Caropreso, S. Matwin, and F. Sebastiani, "A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization," pp. 78-102, Idea Group Publishing, 2001.

[4] Huan Liu and Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Trans. Vol. 17, No. 4, 2005.

[5] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification", IEEE Trans. VOL. 18, NO. 11, 2006.

