ISSN:0975-9646

Distributional Features for Text Categorization Based on Weight

CH.GOWTHAMI, P.RAJA SEKHAR
Department of CSE, Avanthi College of Engg & Tech, Tamaram, Visakhapatnam, A.P., India.

***Abstract*— Text categorization is the task of assigning predefined categories to natural language text. With the widely used "bag-of-word" 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. This paper explores the effect of other types of values, which express the distribution of a word in the document. These novel values assigned to a word are called distributional features, which include the compactness of the appearances of the word and the position of the first appearance of the word. The proposed distributional features are exploited by a tfidf style equation, and different features are combined using ensemble learning techniques. Experiments show that the distributional features are useful for text categorization. In contrast to using the traditional term frequency values solely, including the distributional features requires only a little additional cost, while the categorization performance can be significantly improved. Further analysis shows that the distributional features are especially useful when documents are long and the writing style is casual.**

Index Terms— Text categorization, text mining, machine learning, distributional feature, tfidi.

1. INTRODUCTION

N the last 10 years, content-based document management tasks have gained a prominent status in the information system field, due to the increased availability of documents in digital form and the ensuring need to access them in flexible ways [30]. Among such tasks, Text Categorization assigns predefined categories to natural language text according to its content. Text categorization has attracted more and more attention from researchers due to its wide applicability. Considering the following example, "Here you are" and "You are here" are two sentences corresponding to the same vector using the frequency-related values, but their meanings are totally different. Although this is a somewhat extreme example, it clearly illustrates that besides the appearance and the frequency of appearances of a word, the distribution of a word is also important. Therefore, this paper attempts to design some distributional features to measure the characteristics of a word's distribution in a document. The first consideration is the compactness of the appearances of a word. Here, the compactness measures whether the appearances of a word concentrate in a specific part of a document or spread over the whole document. In the former situation, the word is considered as compact, while in the latter situation, the word is considered as less compact. This consideration is motivated by the following facts. A document usually contains several parts. If the appearances of a word are less compact, the word is more likely to appear in different parts and more likely to be related to the theme of the document.

The contribution of this paper is the following:

1) Distributional features for text categorization are designed. Using these features can help improve the performance, while requiring only a little additional cost.

2) How to use the distributional features is answered. Combining traditional term frequency with the distributional features results in improved performance.

3) The factors affecting the performance of the distributional features are discussed.

4) The benefit of the distributional features is closely related to the length of documents in a corpus and the writing style of documents.

2.HOW TO EXTRACT DISTRIBUTIONAL FEATURES

Recall that the definitions of the two proposed distributional features are both based on the analysis of a word's distribution; thus, modelling a word's distribution becomes the prerequisite for extracting the required features.

2.1. Modeling a Word's Distribution

In this paper, a word's distribution is modeled by two steps: First, a document is divided into several parts; then, the

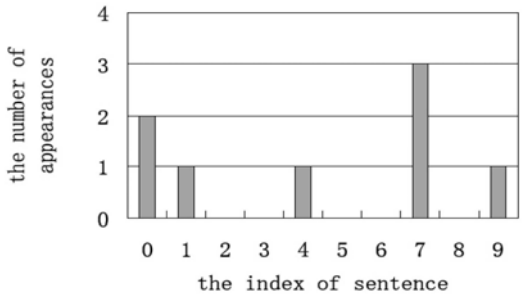


Fig. 1. The distribution of "corn."

Distribution of a word is modeled as an array where each element records the number of appearances of this word in the corresponding part. The length of this array is the total number of the parts.

Now, an example is given. For a document with 10 sentences, the distribution of the word "corn" is depicted in Fig. 1; then the distributional array for "corn" is [2, 1, 0, 0, 1, 0, 0, 3, 0, 1].

2.2 Extracting Distributional Features

Given a word's distribution, this section concentrated on implementing the two intuitively proposed distributional features.

For the compactness of the appearances of a word, three

Implementations are shown as follows (note that under the word distribution model mentioned above, the position of a word's appearance is just the index of the corresponding part):

ComPactPartNum .The number of parts where a word appears can be used to measure the concept of compactness. This is a natural implementation of the idea proposed in the introduction part. As what is mentioned, a word is less compact if it appears in different parts of a document.

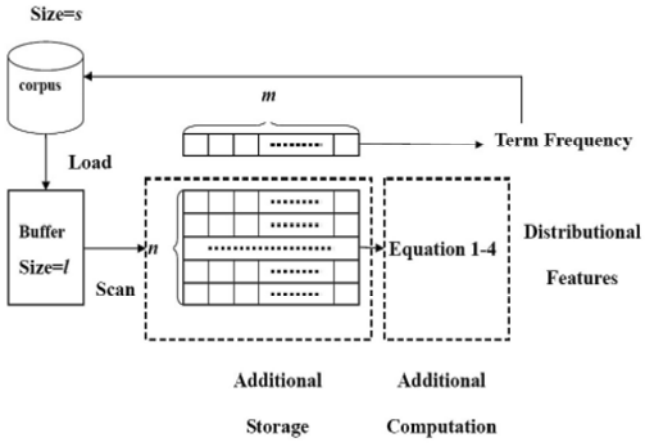


Fig. 2. The process of extracting the term frequency and distributional features

3.HOW TO UTILIZE DISTRIBUTIONAL FEATURES

The term frequency in tf idf can be regarded as a value that measures the importance of a word in a document. As

Name	Function	FA features
GlobalInverse	$f(p, len(d)) = \frac{1}{p+1}$	FA_{GI}
GlobalLogInverse	$f(p, len(d)) = \frac{1}{log(p+2)}$	FA_{GLI}
LocalLinear	$f(p, len(d)) = \frac{len(d)-p}{len(d)}$	FA_{LL}
LocalVLinear	$f(p, len(d)) = \frac{\lvert p-\frac{len(d)-1}{2}\rvert+1}{len(d)}$	FA_{LVL}

TABLE1 - Weighting Functions

discussed , the importance of a word can be measured not only by its term frequency but also by the compactness of its appearances and the position of its first appearance. Therefore, the standard tfidf equation can be generalized as follows:

tfidf (t,d)=importance(t,d)*idf(t)

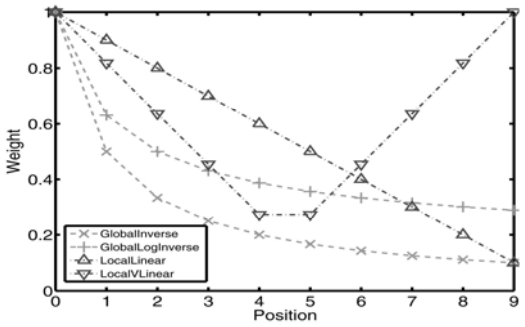


Fig. 3. The trends for different weighting functions.

normalized position. The first three functions assume that the importance decreases with the increase of position, while the last function, LocalVLinear, assumes that the beginning and the end of a document have more importance than the body. Fig. 3 shows the trends of these four functions in a document with 10 parts. Note that in this figure, for each function, the weight is normalized by its maximum weight to facilitate comparison. From this graph, it is clear that LocalVLinear is given such name due to its "V"-like shape.

4. EXPERIMENTS

SVM and kNN are two classifiers that achieved the best performance in a previous comparative study [35]. Thus, in this section, all experiments are based on these two classifiers.

4.1 Data Sets

The Reuters-21578 corpus [19] contains 21,578 articles taken

TABLE 2:

The Contingency Table for Category C_i

Category C_i		Expert Judgement	
		Yes	No
Classifier Judgement	Yes	TP_i	FP_i
	No	FN_i	TN_i

occur in at least two have at least one document in both the training set and the test set are extracted. After eliminating documents that do not belong to any category, there are 7,770 documents in the training set and 3,019 documents in test set. After stemming and stop-word removal, the vocabulary contains 12,158 distinct words that documents of the corpus.

4.2 Performance Measure and Experimental Configuration.

For evaluating the performance on these three corpora, the standard precision, recall, and F1 measure is used. Given the contingency table of category C_i (Table 2), the precision(p_i), recall r_i, and F1 measure F1 of category C_i are calculated as follows:

Pi=TPi/TPi+FPi,ri=TP1/TPi+FNi,F1=2pi/(pi+ri)

These measures can be aggregated over all categories in two ways. One is to average each category's precision, recall, and F1 to get the global precision,

recall, and F1. This method is called macro averaging. The other is based on the global contingency table (Table 3), which is called micro-averaging.

The summarization of the reported combination:

Group	Number of Reported Combinations	Number of Possible Combinations
TF	1: 1 (TF)	1
CP	4: 3 (3 CP features)+1 (CP(best))	$2^3-1=7$
FA	5: 4 (4 FA features)+1 (FA(best))	$2^4-1=15$
TF+CP	4: 3 (3 combinations of TF and each CP feature)+1 (TF+CP(best))	$2^3-1=7$
TF+FA	5: 4 (4 combinations of TF and each FA feature)+1 (TF+FA(best))	$2^4-1=15$
CP+FA	13: 12 (3 x 4 combinations of one CP feature and one FA feature)+1 (CP+FA(best))	$7\times15=105$
TF+CP+FA	13: 12 (3 x 4 combinations of TF, one CP feature and one FA feature)+1 (TF+CP+FA(best))	$7\times15=105$

Parameters are optimized for TF ("bag-of-word" baseline) according to miF1 value. Then, this set of parameters is used for the distributional features.

4.3Effect of Distributional Features

The experiments in this section are designed to explore the effect of the distributional features. The question that we attempt to answer is: are the distributional features useful for text categorization? For eight features (TF+3 CP features+4 FA features).

These features are organized into seven groups: TF, CP, FA, TF + CP, TF + FA, CP + FA and TF + CP + FA. For example, all possible combinations of features from CP and features from FA form the group CP + FA. Due to the limit of the length, the results are reported for a part of combinations of each group, which is summarized in Table 4. Note that TF is the "bag-of- word" baseline.

For other features, the gain of performance compared to the baseline is reported. Suppose the performance of the ith feature (fea_i) and the baseline is $pf(fea_i)$ and $pf(base)$ respectively, the gain (Gain) of fea_i is calculated as follows:

$$Gain(fea_i)\text{-}pf(base)/pf(base)(100\%)$$

Candidate	Average Rank
TF+CP_{PN} +FA_{GI}	4.2
TF+CP_{PN} +FA_{GLI}	5.3
TF+CP_{PN} +FA_{LL}	5.9
TF+CP_{PN} +FA_{LVL}	9.7
TF+CP_{FLD} +FA_{GI}	6.7
TF+CP_{FLD} +FA_{GLI}	7.3
TF+CP_{FLD} +FA_{LL}	6.8
TF+CP_{FLD} +FA_{LVL}	10.8
TF+CP_{PV} +FA_{GI}	**2.5**
TF+CP_{PV} +FA_{GLI}	5.0
TF+CP_{PV} +FA_{LL}	4.7
TF+CP_{PV} +FA_{LVL}	9.3

Average Rank of Different Candidates

The smaller the rank is, the better the performance is.In Table, it is shown that TF + CP_{PV} + FA_{GI} perform the best. In order to show the gap between the selected group of features, i.e., TF, CP_{PV} , and FA_{GI} , and the possible best performance, we also extract the results of different combinations of TF, CP_{PV} , and FA_{GI} from below table and list them in results of distributional features to facilitate comparison

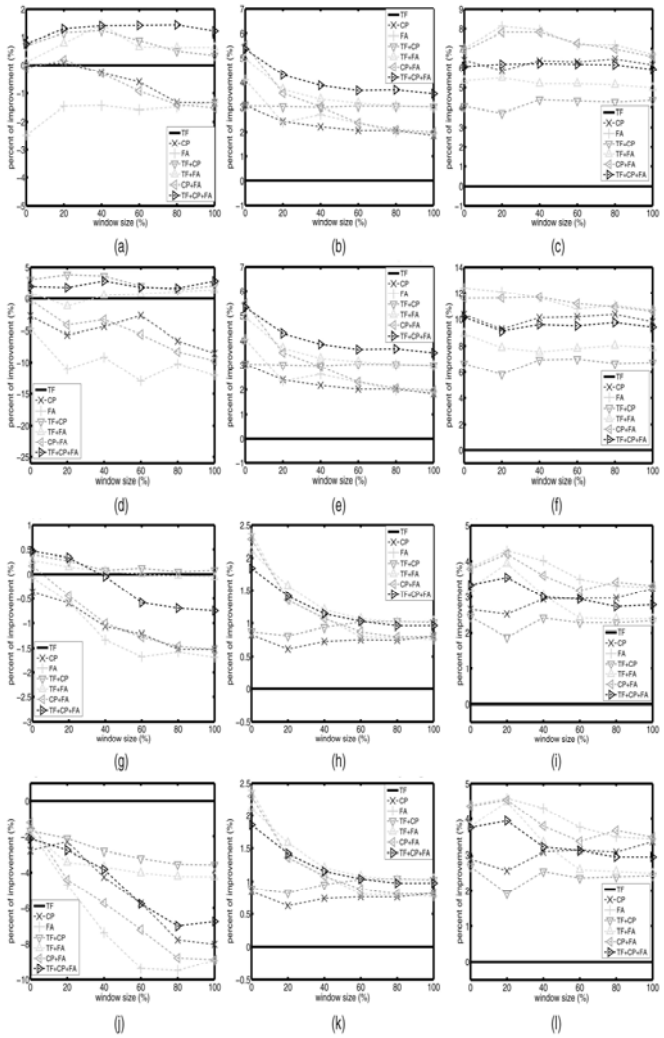
4.4 Factors Influencing the Performance of Distributional Features

As observed, when the distributional features are introduced, there is no obvious improvement on Reuters but a significant improvement on 20 Newsgroup and WebKB. Recall that when the compactness of the appearances of word is introduced, it is assumed that a document contains several parts and the word that only appears in one part is not closely related to the theme of the document. Also, when the position of the first appearance of a word is introduced, it is assumed that the word mentioned late by the author is not closely related to the theme of the document. Intuitively, these two assumptions are more likely to be satisfied when a document contains some loosely related content. After reporting the results of the distributional features using the discourse passage, the window-passage-based distributional features are also tried. For each data set, the maximum length among the 80 percent shorter documents is extracted. Then,five window sizes are tried, from 20 percent to 100 percent of this maximum length, with a gap of 20 percent.[11] The influence of different passages on the performance of the distributional features is shown in Fig. 4. In these figures, the y-axis is the percentage improvement over TF, and the x-axis is the window size (in percentage of the extracted maximum length). The performance of the discourse passage is plotted as the point corresponding to the window size of 0 percent. In these graphs, "CP" corresponds to CP_{PV}, and "FA" corresponds to FA_{GI}.

The first exploration is about the length of a document.

This exploration is based on human's habit of writing. When the length of a document is limited, the author will concentrate on the most related content, such as when writing the abstract of a paper. When there is no limit for the length, the author may write some indirectly related content, such as when writing the body of a paper. The mean length of documents of the three data sets used is reported. Here, the length of a document is measured by its number of words. The average length of a document is 67.9, 115.9 and151.7 respectively, for Reuters, 20 Newsgroup, and WebKB. It seems that the improvement brought by the distributional features is closely related to the mean length of documents. In order to further verify this idea, each of these three data sets is split into two new data sets, i.e., the Short data set and the Long data set, according to the length

of documents. For each data set, the Short data set contains documents with length no more than 100, and the Long data set contains documents with length more than 100. Experiments are repeated for these six new generated data sets discourse-passage-based distributional features.



(l) maF1 Comparison of the distributional features using the discourse passage and the window passages with different sizes. The x-axis denotes the window size (in percentage) of the window passage. The zero position on the x-axis corresponds to the discourse passage. The y-axis denotes the performance improvement (in percentage) over TF. (a) miF1 of kNN on Reuters. (b) miF1 of kNN on 20 Newsgroup. (c) miF1 of kNN on WebKB. (d) maF1 of kNN on Reuters. (e) maF1 of kNN on 20 Newsgroup. (f) maF1 of kNN on WebKB. (g) miF1 of SVM on Reuters. (h) miF1 of SVM 20 Newsgroup. (i) miF1 of SVM on WebKB. (j) maF1 of SVM on Reuters. (k) maF1 of SVM on 20 Newsgroup of SVM on WebKB

According to Table , the distributional features brought more significant improvement on the Long data set than on the Short data set, although there were some exceptions indicated by " " in Table . It seems that the exceptions concentrate on the Reuters data set. We notice that there is a big gap between the baseline of the Short part and the baseline of the Long part on the Reuters data set. In this situation, comparing RGain on the Short and Long parts cannot reflect the effect of the distributional features categorization tasks on Short and Long parts differs significantly.

	*k*NN						SVM					
	Reuters		Newsgroup		WebKB		Reuters		Newsgroup		WebKB	
Gain(%)	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
TF	0.850	0.495	0.816	0.823	0.793	0.659	0.909	0.523	0.877	0.879	0.917	0.898
CP_{PV}	0.7	2.0	3.7**	3.5**	3.2**	2.8*	-0.3	-2.4	1.0**	1.0**	0.8**	1.1
FA_{GI}	-2.6††	-7.2††	4.8**	4.7**	2.2**	3.3*	-0.1	-3.7	2.5**	2.6**	0.3	0.6
TF+CP_{PV}	1.8**	5.5	3.6**	3.4**	2.1**	0.9*	0.2	-1.4	0.9**	0.9**	1.0**	1.5**
TF+FA_{GI}	0.8*	1.0	5.9**	5.8**	2.2**	3.1*	0.5**	-1.6	2.3**	2.4**	1.3**	1.6**
CP_{PV}+FA_{GI}	1.1**	-0.2	6.8**	6.6**	3.1**	2.8	0.1	-2.8	2.5**	2.5**	1.1**	1.7**
TF+CP_{PV}+FA_{GI}	2.0**	3.6**	6.6**	6.4**	2.4**	1.0*	0.3	-2.8	2.0**	2.1**	1.0**	1.5**

Results of the Distributional Features on Three Short Data Sets (Discourse Passage)

The Influence of Document Length on the Results of the Distributional Features Reporting Relative Gain (Discourse Page)

	Reuters				Newsgroup				WebKB			
	miF1		maF1		miF1		maF1		miF1		maF1	
RGain(%)	S	L	S	L	S	L	S	L	S	L	S	L
*k*NN												
TF	0.85	0.63	0.49	0.40	0.82	0.87	0.82	0.85	0.79	0.75	0.66	0.71
CP_{PV}	3.9	-1.1	2.0	-4.4	16.6	20.6	16.5	21.8	12.2	23.6	5.4	26.6
FA_{GI}	-2.6	-2.1	-7.2	-12.9	21.2	25.9	22.0	24.5	8.5	40.1	6.3	43.9
TF+CP_{PV}	10.5	2.8	5.4	-0.4	15.8	21.9	15.6	21.6	8.1	14.9	1.7	16.5
TF+FA_{GI}	4.7	-0.4	1.0	-4.0	26.1	30.5	27.0	29.3	8.5	25.1	6.1	27.9
CP_{PV}+FA_{GI}	6.0	4.1	-0.2	-3.4	30.1	37.9	30.9	37.8	12.0	34.8	5.5	38.2
TF+CP_{PV}+FA_{GI}	11.5	0.6	3.5	-3.9	29.2	36.9	29.9	36.4	9.4	26.7	1.9	30.0
SVM												
TF	0.91	0.73	0.52	0.36	0.88	0.91	0.88	0.90	0.92	0.86	0.90	0.86
CP_{PV}	-0.3	-0.4	-2.4	2.3	6.8	9.6	7.3	8.3	8.6	27.7	9.7	26.9
FA_{GI}	-0.1	2.9	-3.7	-0.8	17.6	29.0	18.7	25.6	3.2	37.6	4.9	38.5
TF+CP_{PV}	2.4	8.3	-1.4	5.8	6.4	12.9	6.8	13.0	11.4	20.7	13.1	20.5
TF+FA_{GI}	5.2	3.2	-1.6	1.5	16.5	27.2	17.3	24.5	14.1	29.2	14.1	29.7
CP_{PV}+FA_{GI}	0.8	6.3	-2.8	3.5	17.6	27.9	18.4	25.1	12.4	38.7	15.2	39.2
TF+CP_{PV}+FA_{GI}	2.7	5.7	-2.8	4.1	14.5	24.4	15.1	23.4	10.8	30.3	11.9	30.6

Baselines on the Short and Long parts are comparable; thus, the comparisons on these two data sets are more convincing. Below Fig. shows that on Reuters, the distribution of the topical words is uniform, while on 20 Newsgroup and WebKB, the topical words are more likely to appear at the beginning of a document. These differences partly explain

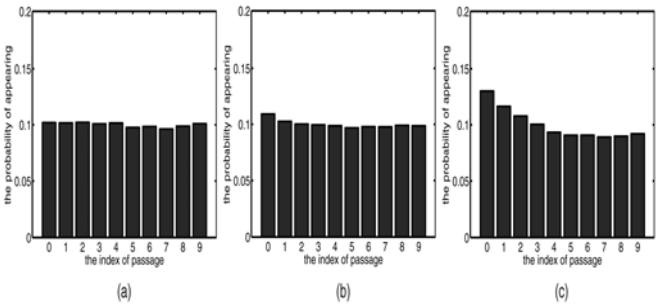


Fig. The average distribution of the topical words for three data sets. (a) Reuters ðstd ¼ 0:0021Þ. (b) 20 Newsgroup (std =0:0035). (c)WebKB (std=0.0137)

	*k*NN						SVM					
	Reuters		Newsgroup		WebKB		Reuters		Newsgroup		WebKB	
Gain(%)	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1	miF1	maF1
TF	0.822	0.550	0.859	0.859	0.788	0.729	0.883	0.554	0.901	0.899	0.901	0.892
CP_{PV}	0.0	-2.7	3.1**	3.0**	6.4**	10.4**	-0.4	-2.8	0.8**	0.8**	2.6**	2.9**
WET_{GI}	-1.9††	-4.4††	1.6**	1.6**	1.4**	4.7	-0.2	0.1	1.3**	1.3**	0.6*	0.7**
TF+CP_{PV}	0.7	3.0	3.0**	3.0**	4.0**	6.6**	0.4*	-1.7	0.9**	0.9**	2.4**	2.7**
TF+WET_{GI}	-0.3†	-1.3	2.6**	2.6**	1.2**	2.6**	0.3	2.4	1.3**	1.3**	0.9**	1.1**
CP_{PV}+WET_{GI}	0.7	2.8	4.4**	4.4**	5.2**	9.2**	0.4*	-2.1	2.0**	2.0**	3.1**	3.6**
TF+CP_{PV}+WET_{GI}	0.8	1.6	4.0**	4.0**	3.7**	6.6**	0.7**	-0.7*	1.5**	1.5**	2.3**	2.6**

4.5 Further Analysis of the FA Features

Since the FA features proposed in this paper consist of two parts: the weighting function f and the strategy of only considering the first appearance of a word, it is necessary to further analyse which part brings the effect of FA features. In order to separate the influence of the weighting function, a group of weighted term frequency (WET) features are generated by using the weighting function f to weight each appearance of a word. Below table shows that FA performs better than wet, especially on 20 Newsgroup and WebKB. The cases where FA performs worse than WET are indicated by "-" Since WET still improves the baseline, it is believed that the effect of FA on 20 Newsgroup and WebKB is brought by both the weighting function and the aggressive strategy that throws all appearances of a word except the first one.

5. CONCLUSION

Previous researches on text categorization usually use the appearance or the frequency of appearance to characterize a word. These features are not enough for fully capturing the information contained in a document. The research reported here extends a preliminary research [33] that advocates using distributional features of a word in text categorization. The distributional features encode a word's distribution from some aspects. In detail, the compactness of the appearances of a word and the position of the first appearance of a word are used. Three types of compactness-based features and four position-of-the-first-appearance-based features are implemented to reflect the different considerations. A tfidf style is constructed, and the ensemble learning technique is used to utilize the distributional features.

Gain(%)	Reuters					Newsgroup				WebKB			
	miF1		maF1			miF1		maF1		miF1		maF1	
X	WET	FA	WET	FA		WET	FA	WET	FA	WET	FA	WET	FA
kNN													
X_{GI}	-1.9	-2.5	-4.4	-4.7		1.6	4.1	1.6	4.1	1.4	7.0	4.7	12.4
$TF+X_{GI}$	-0.3	0.1	-1.3	0.9		2.6	5.0	2.6	4.9	1.2	5.3	2.6	8.9
$CP_{PV}+X_{GI}$	0.7	-0.1	2.8	-0.2		4.4	5.5	4.4	5.5	5.2	6.9	9.2	11.6
$TF+CP_{PV}+X_{GI}$	0.8	0.8	1.6	1.9		4.0	5.4	4.0	5.3	3.7	6.1	6.6	10.1
SVM													
X_{GI}	-0.2	-0.1	0.1	-1.7		1.3	2.4	1.3	2.4	0.6	3.8	0.7	4.4
$TF+X_{GI}$	0.3	0.3	2.4	-1.9		1.3	2.1	1.3	2.1	0.9	3.2	1.1	3.8
$CP_{PV}+X_{GI}$	0.4	0.2	-2.1	-1.2		2.0	2.3	2.0	2.3	3.1	3.8	3.6	4.4
$TF+CP_{PV}+X_{GI}$	0.7	0.5	-0.7	-2.1		1.5	1.8	1.5	1.9	2.3	3.3	2.6	3.8

The comparisions between the FA Feature and the WET Feature with Discourse Passage Reporting Gain Frequency or combined together. Further analysis reveals that the effect of the distributional features is obvious when the documents are long and when the writing style is informal.

ACKNOWLEDGMENT

The authors want to thank the anonymous reviewers for the helpful comments and suggestions. This research was supported by the National Science Foundation of China (under Grants 60505013, 60635030, and 60721002), the Jiangsu Science Foundation (under Grant BK2008018), and the National High Technology Research and Development Program of China (under Grant 2007AA01Z169).

REFERENCES

[1] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. ACM SIGIR '98, pp. 96-103, 1998.

[2] R. Bekkerman, R. Elaine, N. Tishby, and Y.Winter, "Distributional Word Clusters versus Words for Text Categorization,"J. Machine Learning Research, vol. 3, pp. 1182-1208, 2003.

[3] D.CAI, S.P. Yu, J.R. Wen, and WY. Ma, "VIPS: A Vision-Based Page Segmentation Algorithm" Technical Report MSR-TR-2003-79, Microsoft Seattle, Washington, 2003.

[4] J.P. Calan, "Passage Retrieval Evidence in Document Retrieval,"Proc. ACM SIGIR '94, pp. 30310, 1994.

[5] M.F. Caropreso, S. Matwin, and F. Sebastian, "A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization," Text Databases and Document Management: Theory and Practice, A.G. Chin, ed., pp. 78-102, Idea Group Publishing, 2001.

MS.Ch.Gowthami received the B.Tech degree from the Department of Information Technology, A.I.E.T, JNTUniversity, Kakinada in 2009 and She is currently pursuing M.Tech in the Department Of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vishakhapatnam, JNTUniversity. Her research interests include DataMining And Association Rules.

Mr. P.Rajsekhar M.Tech degree from the Department of Computer Science and technology,G.I.T.MUniversityVishakhapatnam in 2009 and His research interests include DataMining And Association Rules.