An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm

Tian Xia
Shanghai Second Polytechnic University, Shanghai, China
Email: xiatian@it.sspu.cn

Yanmei Chai
Central University of Finance and Economics, Beijing, China
Email: chai-4@163.com

***Abstract*—In the process of document formalization, term weight algorithm plays an important role. It greatly interferes the precision and recall results of the natural language processing (NLP) systems. Currently, TF-IDF term weight algorithm is widely applied into language models to build NLP Systems.**

Since term frequency is not the only discriminator which is necessary to be considered in term weighting and make each weight suitable to indicate the term's importance, we are motivated to investigate other statistical characteristics of terms and found an important discriminator: term distribution. Furthermore, we found that, in a single document, a term with higher frequency and close to hypo-dispersion distribution usually contains much semantic information and should be given higher weight. One the other hand, in a document collection, the term with higher frequency and hypo-dispersion distribution usually contains less information.

Based on this hypothesis, by leveraging the Pearson Chi-square Test Statistic, a Term Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm are put forward respectively in this paper. Also, the experiment results at the end of this paper approve the reliability and efficiency of the algorithms.

***Index Terms*—TF, IDF, Term Weight, Natural Language Processing**

I. INTRODUCTION

Document formalization founds the bases of language models, such as Vector Space Model(VSM), Latent Semantic Analysis model, etc. It also consequently impacts on the accuracy of application technologies of Natural Language Processing, such as Information Retrieval, Text Categorization and so on.

In the process of document formalization, documents are represented by document vectors which are expected to indicate as much information of the documents as possible. To make the representation accurate, term weight algorithm plays an important role in the process.

This research is sponsored by Shanghai Municipal Education Commission under Knowledge Innovation Project and Education Highland Building Project - Computer Science and Technology.

TF-IDF is the most widely used term weight algorithm nowadays. However, it has the following drawbacks as well.

TF-IDF is one of the most commonly used term weighting algorithms in today's information retrieval systems. Two parts of the weighting were proposed by Gerard Salton[1] and Karen Spärck Jones[2] respectively. TF, the term frequency, also called Local Term Weight, is defined as the number of times a term in question occurs in a document. Obviously, it is "single document wide" measurement. Since term frequency is an unstable term attribute, TF algorithm is often metamorphosed mathematically and many TF algorithms appear. However, all TF algorithms are only positive correlated to their frequency. IDF, the inverse document frequency, also called Global Term Weight, is based on counting the number of documents in the collection being searched that are indexed by the term. Apparently, it is "documents collection wide" measurement. The product of TF and IDF, known as TF-IDF, is used as an indicator of the importance of a term in representing a document.

However, term frequency is not the only discriminator which is necessary to be considered when calculating the term weight and make it suitable to indicate term importance. Therefore, we are motivated to investigate other statistical characteristics of terms and found an important discriminator after analyzing the distribution data of term statistically.

For the local term weight which is measured in a single document like TF, it is found that a term with higher frequency and close to hypo-dispersion distribution should be given higher weight than one with lower frequency and close to intensive distribution.

On the other hand, for the global term weight which is valued in whole collection of documents like IDF, it is also found that, in such collection, the term with higher frequency and hypo-dispersion distribution usually contains less information.

In this paper, as an improvement to TF-IDF, Term Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm is presented. The former improves TF and is based on each term statistical distribution in a single document. The latter improves

doi:10.4304/jsw.6.3.413-420

IDF and depends on each term distribution in whole collection of documents.

In the end of this paper, an LSA(Latent Semantic Analysis) based information retrieval system and text classifier system are focused as examples for the algorithms' application.

II. PRELIMINARIES

A. Traditional TF-IDF

As defined, TF is the term frequency in a single document. Terms can be words, phases. For documents, the frequency for each term may vary greatly. Therefore, frequency is an important attribute of term to discriminate itself from other terms. Sometimes, term frequency is directly used as the value of TF. That is, the TF value of term i is

$$TF_i = tf_{ik}.$$

where tf_i denotes the frequency of term i in document j.

Since the number of term frequency may be very large, the following formula is also often used to calculate TF value.

$$TF_i = \log_2(tf_{ij}).$$

As for IDF, various formulas have been proposed. A basic formula was given by Robertson[3]. A later discussion between Spärck Jones[4] and Robertson resulted in the following formula of IDF:

$$IDF_i = \log_2(\frac{N}{n_j}) + 1 = \log_2(N) - \log_2(n_j) + 1$$

where N is the total number of documents in the collection and n_j is the number of documents that contain at least one occurrence of the term i.

Your goal is to simulate the usual appearance of papers in a Journal of the Academy Publisher. We are requesting that you follow these guidelines as closely as possible.

B. Drawbacks of Traditional TF

TF-IDF term weight algorithm is widely applied into language models to build NLP Systems. For instance, in SMART system, vector space model (VSM) of text document is put forward by Salton[5]. In the vector space model, a document is represented by a vector of terms. And a term-by-document matrix is used to represent a collection of documents, where each entry represents the weight of a term in a document and is calculated usually via TF-IDF. In addition, in Latent Semantic Indexing[6], the matrix constructed by TF-IDF is usually sparse and factored into product of three matrices using the singular value decomposition since every word does not normally appear in each document.

However, since TF-IDF only takes term frequency into consideration, it also has the following drawbacks.

First, TF algorithm calculates term weight only based on their frequency. That is, term weight is positive correlated to their frequency. Actually, term with higher frequency may be only intensively distributed in a part of the document. Such terms are inclined to represent the content of the part instead of the whole document. However, TF algorithm will assign a higher term weight to such terms. Obviously, it is insufficient to only consider term frequency when calculating its weight.

Second, the intuitive meaning of IDF algorithm is that terms which rarely occur over a collection of documents are valuable. The importance of each term is assumed to be inversely proportional to the number of documents that the term occurs. However, obviously, the term which occurs widely in the document collection but intensively appears in a few documents much probably represents the topic of a document category and is significant for text classifying. However, such scenario is absolutely overlooked by IDF. IDF algorithm will assign a low term weight to such terms. Obviously, it is insufficient to only consider term frequency when measuring its weight.

Third, empty terms and function terms, including conjunctive, preposition, some adverbs, auxiliary term, modal particles, are usually existed with high frequency. This leads to inaccurate weight assignments to such terms. Although stop terms table is always used, this issue cannot be completely resolved.

C. Some TF-IDF improvements

The improvement of TF-IDF mainly concentrated on two topics: taking additional term statistical information into consideration and introducing additional techniques into this field.

For the first topic, position, HTML tags[7] and length of the term have been collected and used into the algorithm. For the other topic, the techniques, such as Mutual information[8], the weight of evidence for text, Information Gain[9], Expected cross entropy[10], etc, have been applied into term weighting.

However, since the formulas of the methods do not process any word distribution information, the first issue discussed above cannot be resolved.

III. AN TF IMPORVEMENT: TERM DISTRIBUTION BASED LOCAL TERM WEIGHT ALGORITHM

A. The Correlation of Term Distribution and Weight in a single document

Please first take the following news for example which appears at JULY 2, 2010, 7:50 A.M with title "Bank Of China To Raise Up To CNY60 billion In Shanghai, HK Rights Issue":

"HONG KONG (Dow Jones)--*Bank of China* Ltd. (3988.HK) said Friday it will *raise* up to CNY60 *billion* from a rights issue in Shanghai and Hong Kong, in an unexpected move aimed at strengthening its capital base after an explosion in lending last year.

The announcement comes just a month after *Bank of China*, one of the country's Big Four state-run lenders, *raised* CNY40 *billion* by selling bonds convertible into its Shanghai-traded shares. Chairman Xiao Gang had earlier said *Bank of China* wouldn't need to *raise* fresh

funds on the mainland stock market after completing the debt sale.

Bank of China, the biggest issuer of new loans in the government-led credit boom last year, said in a statement it expects to issue up to 1.1 rights shares for every 10 existing Shanghai-listed A and Hong Kong-listed H Shares."

Please note that the phrase "Bank of China", an important phrase, occurs in all paragraphs and twice in the second paragraph, that is, the longest one. Also, other important word "raise", "billion", etc usually appear in the longest two paragraphs.

In addition, for the unimportant words, such as "Shares", they intensively appear in the last paragraph. Therefore, they indicate the main topic of part of the document instead of whole document. Also, for other unimportant words, such as "lenders", "fresh", "state-run", etc, they appear only once.

Therefore, we come up to a hypothesis that a term that is only intensively distributed in a part of the document is not very important and should be given lower term weight, because such terms are inclined to represent the content of the part instead of the whole document. In addition, on the contrary, a term that is uniform distributed and widely appeared in the whole document should be given higher weight. That is, the more uniform distribution and wide occurrence of the term, the higher weight is given to it.

To come to this hypothesis, we did the following experiment.

1362 documents are selected from 10 categories, including economy, sports, politics, military, arts, agriculture, industry, life, traffic and culture. For each document, the value of weight W ($W \in (0,2)$) and distribution D ($D = 0,1,2,3,4,5$) are assigned to each term manually by 10 different person. The higher value of distribution D, the more uniform distribution the term is close to and the more widely the term spreads. The average value of W and D is the final manually weight assignment of each term in certain document. The typical values of W and D of terms in single document turn out to be the following figure. Please note that stop words are removed from the document after tagging.

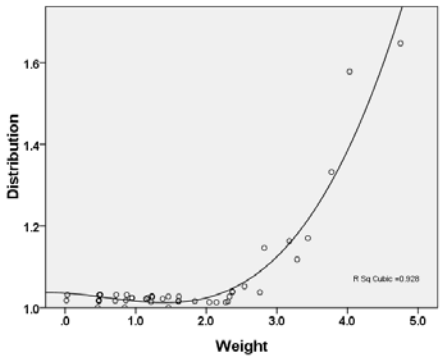

Figure 1. Value of weight W and distribution D

The figure above shows: first, generally the weight of term is positive correlated to but not linear with its uniform distribution extent. Second, the majority of terms are less important with low weights and such term distributes intensively.

B. Distribution based Local Term Weight Algorithm

From the analysis result above, the formula of Distribution based of term weight Algorithm consists of two parts, U and S . U represents the extent of the term's uniform distribution and the other S shows the extension the term spread.

1) Uniform Distribution Extent

To measure Uniform Distribution Extent, we leverage χ^2 experiment method in K. Pearson theory.

After preface process, such as tagging, a document should be transformed into words sequence. To calculate the Uniform Distribution Extent of word number j in the sequence, perform the following method.

Suppose a document contains m paragraphs and C_m words. Also, let interval $(C_{i-1}+1, C_i)$ represent the paragraph number i contains words from number $C_{i-1}+1$ to C_i .

Apparently, any word in the word sequence, including word number j , if it is Uniform Distribution, the probability of the fact that word number j exists in paragraph number i is:

$$r_i = \frac{C_i - C_{i-1}}{C_m} \ (i = 1,2,...,m). \qquad (1)$$

From the Pearson Chi-square Test Statistic, now consider

$$\chi^2{}_j = \sum_{i=1}^{m} \frac{(v_i - nr_i)^2}{nr_i}, \qquad (2)$$

where n is the frequency of word number j in the document and v_i is the frequency of word number j in paragraph number i . It can be shown that, since r_i represent the probability of the fact that word number j exists in paragraph number i , nr_i is the frequency of word number j in paragraph number i if it is Uniform Distribution. Therefore, the numerator is the difference between the actual frequency and the Uniform Distribution frequency of word number j in paragraph number i . Therefore, $\chi^2{}_j$ indicates the Uniform Distribution Extent of word number j .

Furthermore, from the equation (2), lower value of $\chi^2{}_j$ indicates word number j to be more close to Uniform Distribution, which is contrary to the correlation of term distribution and weights. The formula of Uniform Distribution Extent should be:

$$U = \frac{1}{1+\sum_{i=1}^{m}\frac{(v_i - nr_i)^2}{nr_i}} \quad (3)$$

2) Spread Extension

To measure the spread extension of word number j mentioned above, the following formula for spread extension is put forward:

$$S = \log_2(1+\frac{p}{P}) \quad (4)$$

where P is the number of the total paragraphs in the document and p is the number of the paragraphs that word number j exists.

3) Distribution based Local Term Weight Algorithm

In order to combine the formulas above and make the calculate term weight adhere to the figure showed in 3.1, many experiments are performed and the final term weight formula is found:

$$W_{d-l} = \log_2(1+U\times S), \quad (5)$$

that is,

$$W_{d-l} = \log_2(1+\frac{\log_2(1+\frac{p}{P})}{1+\sum_{i=1}^{m}\frac{(v_i - nr_i)^2}{nr_i}}). \quad (6)$$

IV. An IDF Imporvement: Term Distribution Based Global Term Weight Algorithm

A. The Correlation of Term Distribution and Weight in documents collection

Considering the following scenario as an example:

A documents collection contains 10 documents and the occurrence of some words in the collection listed as the following table.

TABLE I.
TERM DOCUMENT FREQUENCY IN A DOCUMENT COLLECTION

Document	Term1 Frequency	Term2 Frequency	Term3 Frequency
Doc 1	10	1	1
Doc 2	13	1	2
Doc 3	9	0	1
Doc 4	1	17	2
Doc 5	0	16	1
Doc 6	11	1	1
Doc 7	2	2	1
Doc 8	0	6	3
Doc 9	0	0	1
Doc 10	1	1	0

If IDF algorithm discussed in II.A is used for calculating global term weight, it is obvious that the three terms are unimportant and assigned with low global term weight as 1.51, 1.32 and 1.15. However, the term 1 and term 2 occurs intensively in a few documents. Such terms much possibly indicates the topic of the documents in which they appear. Therefore, the term 1 and term 2 should have higher weight. Also, actually, the terms similar to term 1 and term 2 are very common in documents. They indicates a document category and are very popular used in many documents.

Each term which is intensively distributed in a group of the documents should be given higher term weight, because such terms are inclined to represent the topic of the documents and are important for text classifying. Otherwise, a term that is uniform distributed and widely appeared in the whole document should be given lower weight, because such terms are inclined to be the frequently used words in almost every document and should be unimportant. In a word, the more uniform distribution and wide occurrence of the term, the lower weight is given to it.

To come to this hypothesis, we did the following experiment which is similar to the experiment discussed in III.A.

In the same 1362 documents which are selected from 10 categories mentioned before, each term are assigned weight W ($W \in (0,2)$) and distribution value D' ($D' = 0,1,2,3,4,5$) manually by 10 different person. However, the meaning of distribution value is different. The higher value of distribution D', the more uniform distribution the term is close to and the more widely the term distributes in the documents collection. The average value of W and D' is the final manually weight assignment of each term in certain document. The typical values of W and D' of terms in the documents collection turn out to be the following figure. Please note that stop words are removed from all documents as well.

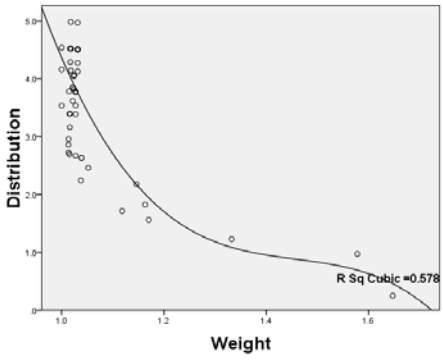

Figure 2. Value of weight W and distribution D'

Obviously, generally the weight of term is negative correlated to but not linear with its uniform distribution extent in documents collection. Also, the majority of terms are lack of semantic information with low weights and such terms distribute uniformly or widely in the collection.

B. Distribution based Global Term Weight Algorithm

From the analysis result above, the formula of Distribution based of Global Term Weight Algorithm consists of two parts, U' and S'. U' represents the extent of the term's uniform distribution and the other S' shows the extension that the term distributes in documents collection.

4) Uniform Distribution Extent

To measure Uniform Distribution Extent, we leverage χ^2 experiment method in K. Pearson theory.

After preface process, each document in documents collection should be transformed into words sequence. To calculate the Uniform Distribution Extent of word number j' in the documents collection, perform the following method.

Suppose a documents collection contains m' documents and C_m' words. The documents are numbered from the first document to the end and the words are numbered from the first word in the first document (document No.1) to the last word in the last document (document No. m').

Also, let interval $(C_{i-1}'+1, C_i')$ represents the document number i contains words from number $C_{i-1}'+1$ to C_i'.

Apparently, any word in the word sequence, including word number j', if it is Uniform Distribution, the probability of the fact that word number j' exists in document number i' is:

$$r_i' = \frac{C_i' - C_{i-1}'}{C_m'} (i = 1, 2, ..., m'). \quad (7)$$

From the Pearson Chi-square Test Statistic, now consider

$$\chi^2{}_j = \sum_{i=1}^{m'} \frac{(v_i' - n' r_i')^2}{n' r_i'}, \quad (8)$$

where n' is the frequency of word number j' in the document collection and v_i' is the frequency of word number j' in document number i'. It can be shown that, since r_i' represent the probability of the fact that word number j' exists in document number i', $n' r_i'$ is the frequency of word number j' in document number i' if it is Uniform Distribution. Therefore, the numerator is the difference between the actual frequency and the Uniform Distribution frequency of word number j' in document number i'. Therefore, $\chi^2{}_j$ indicates the Uniform Distribution Extent of word number j' in the whole document collection.

Furthermore, from the equation (8), lower value of $\chi^2{}_j$ indicates word number j' to be more close to Uniform Distribution, which is consistent with the correlation of term distribution and weights. However, the χ^2 value for each word is quite different from each other. Therefore, the formula of Uniform Distribution Extent should be modified as the following:

$$U' = 1 + \sum_{i=1}^{m'} \frac{(v_i' - n' r_i')^2}{n' r_i'} \quad (9)$$

5) Spread Extension

To measure the spread extension of word number j mentioned above, the following formula for spread extension is put forward:

$$S' = \log_2 (1 + \frac{p'}{P'}) \quad (10)$$

where P' is the number of the total documents in the document and p' is the number of the documents that contains word number j'.

6) Distribution based Global Term Weight Algorithm

In order to combine the formulas above and make the calculate term weight adhere to the figure showed in 3.1, many experiments are performed and the final term weight formula is found:

$$W_{d-g} = \log_2 (1 + U' \times S'), \quad (11)$$

that is,

$$W_{d-g} = \log_2 (1 + (1 + \sum_{i=1}^{m'} \frac{(v_i' - n' r_i')^2}{n' r_i'}) \cdot \log_2 (1 + \frac{p'}{P'})) \quad (12)$$

V. Experiments and Results

To check the efficiency of the weight algorithms, an information retrieval system and a text classifier are developed based on LSA(Latent Semantic Analysis) model. The following tests are performed respectively to evaluate the efficiency of the weight algorithms in both Natural Language Processing application scenarios.

A. IR System Experiments

To compare the weight algorithms, an information retrieval system is developed based on LSA(Latent Semantic Analysis) model and its precision and recall results are utilized for evaluation.

The corpus, collected from portal sites by the VIPS module in the IR system automatically, consist of 10 categories, including economy, sports, politics, military, arts, agriculture, industry, life, traffic, culture and more than 1.5 million documents.

For the LSA module of the IR system, the term-document matrix is constructed by several weight algorithms, such as TF, W_{d-l}, IDF, W_{d-g}, TF-IDF and $W_{d-l} \cdot W_{d-g}$.

The precision and recall results of the IR system by using TF and Distribution based Local Term Weight Algorithm are shown in the following table.

TABLE II.
PRECISION AND RECALL RESULTS OF TF AND W_{d-l} ALGORITHM FOR EACH CATEGORY

Precision &Recall	TF		W_{d-l}	
Dimension reserved	51		62	
	P	***R***	***P***	***R***
economy	0.737	0.690	0.795	0.752
sports	0.770	0.656	0.821	0.697
politics	0.787	0.735	0.767	0.786
military	0.705	0.781	0.788	0.769
arts	0.764	0.729	0.762	0.802
agriculture	0.786	0.707	0.852	0.746
industry	0.723	0.780	0.797	0.790
life	0.717	0.793	0.785	0.889
traffic	0.742	0.742	0.806	0.792
culture	0.798	0.697	0.823	0.785
Average	0.753	0.731	0.799	0.781

The precision and recall results of the IR system by using IDF and Distribution based Global Term Weight Algorithm are shown in the following table.

TABLE III.
PRECISION AND RECALL RESULTS OF IDF AND W_{d-g} ALGORITHM FOR EACH CATEGORY

Precision &Recall	IDF		W_{d-g}	
Dimension	46		52	
	P	***R***	***P***	***R***
economy	0.722	0.731	0.734	0.752
sports	0.757	0.681	0.782	0.680
politics	0.790	0.723	0.794	0.755
military	0.732	0.803	0.728	0.816
arts	0.771	0.729	0.785	0.720
agriculture	0.764	0.718	0.807	0.746
industry	0.682	0.769	0.737	0.770
life	0.708	0.773	0.725	0.806
traffic	0.786	0.732	0.796	0.772
culture	0.802	0.742	0.823	0.755
Average	0.751	0.740	0.771	0.757

The precision and recall results of the IR system by using TF-IDF and the product of Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm are shown in the following table.

Obviously, Distribution based Local Term Weight Algorithm shows much more efficiency than TF weight algorithm. Combined with IDF, it also come up with better precision and recall results than TF-IDF.

The following figure also shows the average precision and recall results for the four weight algorithms in the latent semantic space built by the IR system.

TABLE IV.
PRECISION AND RECALL RESULTS OF TF-IDF AND $W_{d-l} \cdot W_{d-g}$ ALGORITHM FOR EACH CATEGORY

Precision &Recall	TF-IDF		$W_{d-l} \cdot W_{d-g}$	
Dimension	47		55	
	P	***R***	***P***	***R***
economy	0.734	0.763	0.772	0.784
sports	0.791	0.714	0.807	0.787
politics	0.762	0.864	0.797	0.859
military	0.79	0.703	0.821	0.857
arts	0.801	0.831	0.858	0.882
agriculture	0.823	0.835	0.870	0.940
industry	0.885	0.794	0.898	0.901
life	0.822	0.902	0.841	0.909
traffic	0.814	0.793	0.877	0.795
culture	0.801	0.721	0.898	0.829
Average	0.802	0.792	0.844	0.854

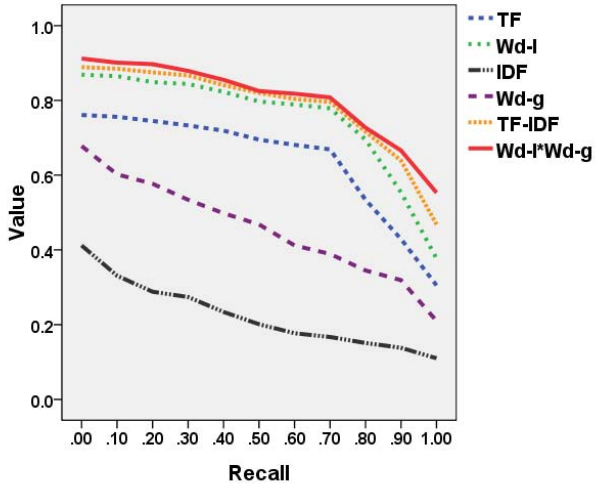

Figure 3. Comparison of the 6 weight algorithms in latent semantic space

From the figure above, for IR issue, IDF itself almost has no effect because it only decreases the weight of terms that widely appear in document collection. TF itself does have some effect. However, when also considering term distribution, the Distribution based Local Term Weight Algorithm shows its efficiency in IR issue. In addition, combined with Distribution based Global Term Weight Algorithm, $W_{d-l} \cdot W_{d-g}$ has the best efficiency among the six term weight algorithms.

B. Text Classifier System Experiments

To evaluate the efficiency of the weight algorithms, we also develop a text classifier system based on LSA model by using the same corpus.

When formalizing documents, each item in the term-document matrix is calculated via the 6 weight

algorithms(TF, W_{d-l} , IDF, W_{d-g} , TF-IDF and $W_{d-l} \cdot W_{d-g}$).

The precision and recall results of the Text Classifier system by using TF and Distribution based Local Term Weight Algorithm are shown in the following table.

TABLE V.
PRECISION AND RECALL RESULTS OF TF AND W_{d-l} ALGORITHM FOR EACH CATEGORY

Precision &Recall	**TF**		W_{d-l}	
Dimension reserved	47		57	
	P	***R***	***P***	***R***
economy	0.722	0.691	0.725	0.701
sports	0.779	0.677	0.785	0.687
politics	0.733	0.735	0.737	0.739
military	0.718	0.634	0.721	0.645
arts	0.731	0.727	0.730	0.735
agriculture	0.777	0.725	0.784	0.698
industry	0.759	0.719	0.769	0.708
life	0.725	0.795	0.756	0.807
traffic	0.733	0.754	0.769	0.766
culture	0.796	0.677	0.818	0.693
Average	0.747	0.713	0.759	0.718

The precision and recall results of the Text Classifier system by using IDF and Distribution based Global Term Weight Algorithm are shown in the following table.

TABLE VI.
PRECISION AND RECALL RESULTS OF IDF AND W_{d-g} ALGORITHM FOR EACH CATEGORY

Precision &Recall	**IDF**		W_{d-g}	
Dimension	53		59	
	P	***R***	***P***	***R***
economy	0.658	0.732	0.728	0.751
sports	0.713	0.769	0.756	0.857
politics	0.742	0.742	0.737	0.769
military	0.719	0.689	0.758	0.725
arts	0.772	0.717	0.798	0.735
agriculture	0.798	0.768	0.847	0.761
industry	0.765	0.743	0.802	0.788
life	0.723	0.790	0.766	0.867
traffic	0.768	0.698	0.819	0.766
culture	0.812	0.768	0.818	0.787
Average	0.747	0.741	0.783	0.780

The precision and recall results of the Text Classifier system by using TF-IDF and the product of Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm are shown in the following table.

TABLE VII.
PRECISION AND RECALL RESULTS OF TF-IDF AND $W_{d-l} \cdot W_{d-g}$ ALGORITHM FOR EACH CATEGORY

Precision &Recall	**TF-IDF**		$W_{d-l} \cdot W_{d-g}$	
Dimension	48		61	
	P	***R***	***P***	***R***
economy	0.713	0.689	0.767	0.706
sports	0.797	0.695	0.841	0.787
politics	0.724	0.783	0.789	0.869
military	0.768	0.812	0.832	0.870
arts	0.791	0.723	0.824	0.821
agriculture	0.825	0.725	0.864	0.768
industry	0.734	0.799	0.876	0.879
life	0.717	0.792	0.786	0.847
traffic	0.746	0.764	0.809	0.826
culture	0.853	0.692	0.898	0.773
Average	0.767	0.747	0.829	0.815

Obviously, in a text classifier system, since W_{D-G} (Distribution based Global Term Weight Algorithm) amplifies the weight of the terms that are centralized distributed in part of the documents collection and should contain much information for a document category, it much fits text classifying and shows much more efficiency than IDF weight algorithm.

Combined with W_{D-L} (Distribution based Local Term Weight Algorithm), the product of W_{D-L} and W_{D-G} also come up with better precision and recall results than TF-IDF.

The following figure also shows the average precision and recall results for the four weight algorithms in the latent semantic space built by the text classifier system.

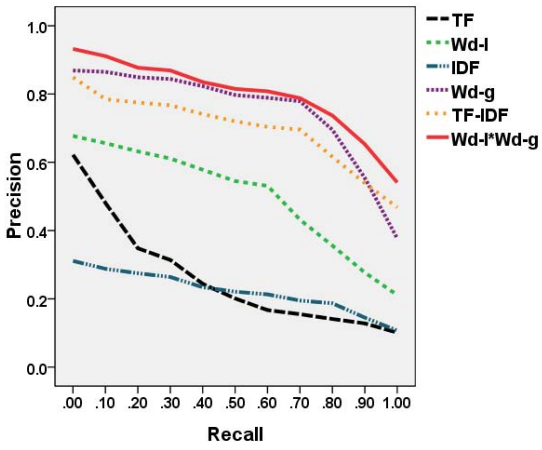

Figure 4. Comparison of the 6 weight algorithms in latent semantic space

Apparently, since IDF does not have any document category information in document collection, IDF itself has no effect for Text Classifying issue. Please check the Distribution based Global Term Weight Algorithm. Since

term distribution information contains document category information, it is much efficient in this issue. Also, Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm $W_{d-l} \cdot W_{d-g}$ has the best efficiency among the six term weight algorithms.

VI. Conclusion

On conclusion, the Distribution based Local Term Weight Algorithm and Global Term Weight Algorithm in this paper improves TF and IDF algorithm respectively through introducing term distribution data into term weighting research.

Moreover, by leveraging the Pearson Chi-square Test Statistic, the both weight algorithms are able to simulate the manual weight assignments.

Finally, the test results of the IR system and Text Classifier system also show their convincible efficiency.

Acknowledgment

We would like to thank for the support provided by Shanghai Municipal Education Commission under Knowledge Innovation Project and Education Highland Building Project - Computer Science and Technology.

References

[1] Salton, G., "Automatic information organization and retrieval", *McGraw-Hill*, New York. 1968

[2] Spärck Jones, K., "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. pp. 28, 11–21, 1972.

[3] Robertson, S., "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, vol. 60, pp. 503–520, 2004.

[4] Spärck Jones, K., "IDF term weighting and IR research lessons", *Journal of Documentation*, vol. 60, pp. 521–523, 2004

[5] Salton, G., "Introduction to modern information retrieval", *Auckland. McGraw-Hill,ew York*, 1983.

[6] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R., "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, pp. 391–407, 1990.

[7] CHU Jian-chong, LIU Pei-yu, WANG Wei-ling, "Improved approach to weighting terms in Web Text", *Computer Engineering and Applications*, vol. 43, pp. 192-194, 2007.

[8] Harksoo Kim, Jungyun Seo, "Cluster-Based FAQ Retrieval Using Latent Term Weights", *Intelligent Systems*, pp. 58-65, April 2008.

[9] LU Song, LI Xiao-li, BAI Shuo, WANG Shi, "An Improved Approach to Weighting Terms in Text", *JOURNAL OF CHINESE INFORMATION PROCESSING*. vol. 14(6), pp. 8-13, 2000.

[10] LIU Yun-feng , QI Huan, Xiang'en Hu, Zhiqiang Cai, "A Modified Weight Function in Latent Semantic Analysis". *JOURNAL OF CHINESE INFORMATION PROCESSING*. vol. 19(6), pp. 64-69, 2005.

[11] Xia Tian, Wang Tong, "An improvement to TF: Term Distribution based Term Weight Algorithm". *Proc. Of IEEE NSWCTC 2010*, April 2010.

Tian Xia was born in Henan Province, China, on April 1, 1979. He received the Ph.D. degrees in computer science and technology from the East China Normal University, Shanghai, China in 2007.

He is currently working as an associate professor at the Shanghai Second Polytechnic University, Shanghai, China. He has published 16 papers in international journals, the national core journals and international conference, in which 9 papers were indexed by SCI/EI. His major research interests include natural language processing, pattern recognition, information retrieval, etc.

Mr. Xia is a member of China Computer Federation and has been involved in many major projects, such as Knowledge Innovation Project and Education Highland Building Project - Computer Science and Technology.

Yanmei Chai received the Ph.D. degrees in computer science and technology from the Northwestern Polytechnic University, Xi'an, China in 2007. She had worked as a postdoctoral researcher at the Tsinghua University, Beijing, China from 2007 to 2009. She is currently working as a lecture at the Central University of Finance and Economics, Beijing, China. She has been involved in many major national projects and transnational horizontal project. She has published 22 papers in the national core journals and international conference, in which 14 papers were indexed by EI/SCI.Her research interests include Image Processing,Pattern Recognition, Information Retrieval and Information Security etc.