Journal of Computer and Communications, **2013, 1, 22-27**
http://dx.doi.org/10.4236/jcc.2013.14004 Published Online October 2013 (http://www.scirp.org/journal/jcc)

Short Text Classification Based on Improved ITC

Liangliang Li, Shouning Qu

School of Information Science and Engineering, University of Jinan, Jinan, China.
Email: liliangliang24@126.com

Received September 9th, 2013; revised October 6th, 2013; accepted October 13th, 2013

ABSTRACT

The long text classification has got great achievements, but short text classification still needs to be perfected. In this paper, at first, we describe why we select the ITC feature selection algorithm not the conventional TFIDF and the superiority of the ITC compared with the TFIDF, then we conclude the flaws of the conventional ITC algorithm, and then we present an improved ITC feature selection algorithm based on the characteristics of short text classification while combining the concepts of the Documents Distribution Entropy with the Position Distribution Weight. The improved ITC algorithm conforms to the actual situation of the short text classification. The experimental results show that the performance based on the new algorithm was much better than that based on the traditional TFIDF and ITC.

Keywords: ITC; Text Classification; Short Text

1. Introduction

Short text classification is that categorizing each short document like the news, BBS text, comments, instant chat (MSN/QQ) in a document collection based on a predefined set of categories [1]. Owing to the characteristics of the short text: text content is less and some of the texts are not in strict text format, the sentence length is various and complex structure. Often some only contain very few words—about two hundred words [2]. As a result, the text feature is not obvious. The performance of the conventional algorithm is not satisfied. At present, we mainly concentrate on the long text in the field of text classification; however the studies on short text are to be mature. With the popularization of Internet, the Internet has become the main channel through which people get the information. For instance, we browse the news on the Internet, we can also share our information with friends on the Social Network Site like Facebook, Twitter, Microblog and so on, and people mainly get touch with the short text in these activities. It could be said that the amount of the short text occupies a large proportion. How to handle information which comes from the short text is more and more important, which will provide a new development platform.

Short text classification as the conventional text classification can be divided into four steps: document preprocess, feature selection, document VSM representation and classifier training [3]. A huge number of training documents and vector dimensions are the big challenge, but these vectors are not all helpful for text classification and this situation may lead to the low efficiency, so we need to preprocess the documents to filter some terms which are little significant for text classification. Filtering one or several feature terms may not influence the performance very much in the long text classification. However if we ignore some feature terms in the short text classification, this may lead to the bad result, so we should guarantee more accuracy when we preprocess these short documents [4]. Feature selection has an important influence in the aspects of calculate time and the performance; hence feature selection is the key of short text classification. The effect of the term frequency is weaker than that in the long text, we must not ignore the influence of some low frequency terms, especially some documents like BBS, these documents are not in strict format and the writing style is casual which may also bring extra burden for short text feature selection objectively. In conclusion, feature selection of the short text selection is more difficult compared to the long text.

The rest of the paper is organized as follows: Section 2 presents the research status and the preliminary concepts, regarding ITC, TFIDF, Information Entropy and Position Distribution Weight. Section 3 presents the details of the improved ITC. Section 4 analyzes the experiments results, and Section 5 concludes and gives the pointer to the future work.

2. Related Work

Research Status. Currently other researchers had make some achievements in the field of short text classification. At the earliest, we treat this similar convenitional long text classification and still adopt a series of traditional feature selection algorithms which are adopted in the field of long text classification as The TF (Term Frequency), DF (Document Frequency), Mutual Information, Information Gain, Odds Ratio, Expected Cross Enropy and TFIDF (Term Frequency inverse Document Frequency), χ^2 statistics (CHI) and so on [5], after that we will select one classifier to train. The forward experiments have indicated that: though this way is simple and rapid, the accuracy of result is low. Liu Wenyin who is from City University of Hong Kong proposed a short text modeling algorithm in which he combined the semantic information with the statistic information to calculate short text similarity [6]. Yu Yong who comes from Shanghai Jiao Tong University proposed a hierarchical keyword extracting algorithm to make a breakthrough in the field of BBS document classification [7]. In addition, Wang Sheng and his colleagues who come from Chong qing University of Posts and Telecommunications have created a new short text classification algorithm based on hyponymy relations [8]. Cui Zhengyan of Henan University put forward a short text classification algorithm based on microblog information [9]. He mapped the feature term to the relative semantic concept while incorporating the Ontology. Han Zhongming of BTBU constructed microblog Emotional Dictionary based on HowNet Emotional Dictionary [10]. After that process, he would construct an automata machine to get the emotion tendentiousness of a text to achieve the target of text classification. Zhang Zhifei of USTC has proposed a text similarity algorithm based on probability of LDA topic model [11]. In addition, some researchers also have make a contribution in the fields of product evaluation and note [12,13]. In a word, we have bear fruit from the point of semantics in the filed of short text classification, nonetheless we have not get big breakthrough in the point of the statistics, and currently researchers often only study one aspect of short documents, for instance, one-side news or BBS or microblog, but no algorithms can be in common use in different fields. In this paper, we aim at the weakness of current short text classification. At first, we summarize the structural features of the short documents, and then we improve the conventional ITC with the Information Entropy and the Position Distribution Weight. At last, this paper proposes a new feature selection algorithm which is universal in the short documents which is from different fields. We hope to make a breakthrough in the statistics point. The experiments show that the recall ratio and precision ration of the improved ITC is a certain higher than the traditional ITC and TFIDF.

TFIDF. The TFIDF feature selection algorithm proposed by Salton is a classical algorithm. After many years, the transforms of the IFIDF is more and more. At present, the universal formula is [14]:

$$w_{id} = \frac{tf_{id} * \log\left(\frac{N}{n_i} + 0.01\right)}{\sqrt{\sum_{i=0}^{n} tf_{id}^2 * \log^2\left(\frac{N}{n_i} + 0.01\right)}} \quad (1)$$

Where the tf_{id} indicates the term frequency of term t_i in the document d, the N is the total number of the documents collection, n_i is the number of documents in the collection that the term t_i occurs in. The main idea of the TFIDF is that if a word or phrase in an article appears in the high frequency but rarely appears in other articles, it indicates that the word or phrase has the very good category differentiate ability.

The Conventional ITC. ITC is an improved algorithm of TFIDF, it adopts the logarithm of term frequency instead of term frequency, thereby this substitution weakens the impact of the term frequency. The ITC algorithm is expressed by the equation as follows [15]:

$$w_{id} = \frac{\log\left(tf_{id}\right) * \log\left(\frac{N}{n_i} + 0.01\right)}{\sqrt{\sum_{i=0}^{n} \log^2\left(tf_{id}\right) * \log^2\left(\frac{N}{n_i} + 0.01\right)}} \quad (2)$$

Explicit the defects of the ITC and TFIDF. The TFIDF feature selection algorithm is of obvious deficiencies [16]. Firstly, it is greatly influenced by the document training collection. If N is high, while n_i is also big, then the IDF values is small in terms of the IDF formula, this indicates that the category ability differentiate of the term or phrase is low. In fact, if a term appears in a category of documents frequently, therefore this term can be a feature of the category of documents; the term should be given higher weight. For example, assuming that there are 30 sports documents and 30 finance and economic documents, the IDF value of the term “stock” is just 0.303 in the finance and economic documents collection, while the IDF value is 1.447 in the sports documents collection. However, it is obvious that stock is one feature of the category of finance and economic documents collection, this term “stock” should be given bigger weight. Because we decided to adopt the ITC algorithm, the ITC will reduce the effect of term frequency, the influence of the IDF flaw is evident, and so we need to improve the traditional ITC. Secondly, the result relies too much on the term frequency. In the field of short text classification, sometimes some low-frequency words are more representative than some high-

frequency words. For instance, some terms in the news documents only appear in the title or the first paragraph, the TFIDF weight is small, but they should be given higher weight according to the actual situation.

Although this ITC algorithm has get over the second flaw of TFIDF and solved the problem of term frequency dependency, still can not overcome the first flaw of TFIDF. Also the ITC only conquers the first flaw in the statistics point. In fact, we absolutely get over this flaw to get better performance in many points in terms of the structural features of the short documents.

Documents Distribution Entropy. We introduce the following definitions before we improve the conventional ITC.

Definition 1. Assuming that there is *n* information which has the same probability, and the probability of each one is 1/*n*., the content every information transforms is [17]:

$$I(x) = \log \frac{1}{p(x)} = -\log p(x) = \log n \tag{3}$$

Definition 2. For any random variable *X*, if the probability $P = (p(x_1), p(x_2), \cdots p(x_3))$, then the information quality this distribution transforms is called the entropy of *P* [18]:

$$I(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{4}$$

If *X* is (0.5, 0.5), then $I(X)$ is 1. If *X* is (0.67, 0.33), then $I(X)$ is 0.92. If *X* is (1, 0), then $I(X)$ is 0. As is shown that, if the probability distribution is more uniform, the information content is bigger.

Definition 3. In a documents collection, the Document Distribution Entropy of a term is expressed by the following equation:

$$DDEC(t_i) = p(t_i) \log p(t_i) = \frac{n_i}{N} \log \frac{n_i}{N} \tag{5}$$

Where n_i is the number of documents in the collection that the term t_i occurs in. The *N* is the total number of the documents collection. Because the value of this algorithm is negative, so we adopt the absolute value of the result. The document distribution probability is higher, the entropy is bigger. This indicates that the document distribution is more uniform, and then the term can be good representative in the category. The algorithm of the Document Distribution Entropy can get over the first flaw of the ITC.

Position Distribution Weight. Through our careful observation of short text structure, we have concluded that some term which on special position can represent the text excellently, but the term frequency of these terms is not the highest. For instance, terms which appear in the title or the first paragraph are often the feature of the news documents. In the same situation, reply topics in the BBS.

Terms Distribution Entropy. In addition, we also can make use of the term distribution deviation to get the term distribution situation. If the term distribution is more uniform and wide range, then the term is more meaningful for the document, so the term should be given higher weight. The Term Distribution Weight formula is:

$$TF(t_i) = TF(t_i) + a \times Ws \tag{6}$$

$$Wpos(t_i) = \frac{\sum_{k=1}^{TF(t_i)} Dev(k)}{TF(t_i)} \tag{7}$$

Where *a* is the coefficient, *Ws* is the weighted value of term t_i when t_i occurs on the special position. In this paper, the important positions include three parts: title, abstract, the first paragraph, and we set different value for different position in terms of different importance. If a term occurs in all the special positions then needs the sum of three position weighted values. $Dev(k)$ is the distribution deviation of the term when it occurs the *K* time. The algorithm process is: at first, we calculate the average distribution position of a term, and then get the position deviation of the term first occurs. At last, we need to accumulate these term distribution deviations; the deviation is the absolutely value of the difference of the term current position and previous position, the average distribution position. According to the formula, if a term occurs uniformly in the documents, then the term will get higher weight. Assuming that a term occurs in the special positions of the document, and even occurs in the document uniformly at the same time, the term will bet assigned for the biggest weight value.

3. Improved ITC

In this paper, we add the factors of the Information Entropy and Position Distribution Weight, the improved ITC algorithm is:

$$w_{id} = ITC \times DDEC(t_i) \times Wpos(t_i) \tag{1}$$

Where $DDEC(t_i)$ is the Information of the term in a category. According to the analysis above, if the term t_i distribution is more uniform, the $DDEC(t_i)$ value is bigger, this circumstance shows that t_i can represent the category better and has the stronger ability of category classification. For this work, we have collected a huge amount of shot documents which from the BBS, news, comments and so on. After our careful observation, we have concluded that these short documents all have a common characteristic that the authors try the best to take the advantage of the shortest sentences to express the main idea, which brings us a congenital advantage

that the possibility of the term trending to be the noise is lower. Comparing with the long text, the term still represents a category even if the term only occurs a time in a document, so the Information Entropy factor have the more active affect in the short text classification. Also thanks to our meticulous, the position factor weighted affect is more obvious in the short documents such as news, BBS and so on. Because the term frequency affect is weakened, so how to balance the affect and make the term which may be the feature but has the low frequency given the weight the term should own? Thus, we introduce the position factor. $Wpos(t_i)$ is the term position factor. According to our concept of the Position Distribution Weight, we give the title position the biggest weight, the abstract position takes the second place, and the first paragraph position weight is the minimum. At the same time, if a term distribution is more uniform in the document, then the term will be given bigger weight. In a word, if a term not only occurs on the important position but also uniformly, then it will be given the maximum weight.

The improved ITC formula is:

$$w_{id} = \frac{\log(tf_{id}+wa)\times\log\left(\frac{n_i}{N}+0.01\right)}{\sqrt{\log^2(tf_{id}+wa)\times\log^2\left(\frac{n_i}{N}+0.01\right)}} \times\left(\frac{n_i}{N}\right)\log\left(\frac{n_i}{N}\right)\times Wpos(t_i) \quad (2)$$

Where tf_{id} is the term frequency in a document, if tf_{id} is higher, the possibility of the term tending to be the feature is higher. *Ws* is the weighted value of important position. *a* is coefficient. *N* is the total number of the training documents collection. n_i is the number of documents in the collection that the term t_i occurs in. Comparing with the noise problem of the long text classification, the short documents have advantages because of its concise and comprehensive style, especially after text preprocess. The noise will not enhance too much with increasing number of the training documents, so the total number of training documents collection is higher, the performance is better. After our improvement, the new algorithm solves the problems of the traditional ITC, and conforms to the situation of the short text classification better.

4. Experimental Evaluation

For the evaluation of the proposed improved ITC algorithm, we adopt the Vector Space Model. Firstly, we calculate the similarity between a document to be classified and the predefined categories, secondly rank the category, lastly make the category which is the highest be the category which the document will belong to. The current similarity formula can be expressed by the cosine measure given in the following equation:

$$sim(d_i, c_j) = \frac{\sum_{k=1}^{n} w_{ik}\times w_{ik}}{\sum_{k=1}^{n} w_{ik}^2 \times w_{ik}^2} \quad (1)$$

Where d_i represents the documents to be classified, and w_{ik} represent the weight of the *K*th feature of the document d_i and the category c_j respectively. For testing the performance of the new algorithm, we have collected 500 documents downloaded from the Internet news sites. These documents involve five categories in all which include education, sports, military, car, finance, average 100 documents among every category. The term number of every document is lower than 200 after preprocess. We choose 300 documents as the training collection randomly, the rest as the test collection.

Let us define two evaluation parameters such as R (recall factor), P (precision factor) as follows [19]:

$$precision = \frac{|\{relevant\}\cap\{retrieved\}|}{|\{relevant\}|} \quad (2)$$

$$recall = \frac{|\{relevant\}\cap\{retrieved\}|}{|\{relevant\}|} \quad (3)$$

$$F-score = \frac{recall\times precision}{recall + precision/2} \quad (4)$$

Where, *Retrieved* is the documents which are retrieved in the documents to be classified, Relevant is the documents which are relevant to the query.

We test the conventional TFIDF, ITC and improved ITC respectively, the result is shown in the following **Tables 1-3** and **Figure 1**.

Table 1. Recall and Precision Rate of The TFIDF.

	Sports	Military	Medicine	Car	Finance
Sports	28	5	3	2	2
Military	1	25	3	7	4
Medicine	2	5	33	0	0
Car	6	3	3	26	2
Finance	3	1	2	4	30
recall	0.7	0.6125	0.8125	0.625	0.75
precision	0.7143	0.76	0.7576	0.6923	0.7333
	Sports	Military	Medicine	Car	Finance
Sports	28	5	3	2	2
Military	1	25	3	7	4
Medicine	2	5	33	0	0
Car	6	3	3	26	2
Finance	3	1	2	4	30
recall	0.7	0.6125	0.8125	0.625	0.75
precision	0.7143	0.76	0.7576	0.6923	0.7333

Table 2. Recall and Precision Rate of the ITC.

	Sports	Military	Medcine	Car	Finance
Sports	31	3	4	2	0
Military	1	26	3	6	4
Medicine	3	3	32	2	2
Car	2	1	5	27	5
Finance	3	2	0	3	32
recall	0.7625	0.625	0.8	0.675	0.8
Precision	0.7857	0.8075	0.7813	0.7778	0.8125

Table 3. Recall and Precision Rate of the Improved ITC.

	Sports	Military	Medicine	Car	Finance
Sports	32	1	3	4	0
Military	0	31	2	5	2
Medicine	2	0	34	3	1
Car	2	3	0	31	4
Finance	1	1	2	1	35
recall	0.8	0.8125	0.825	0.8125	0.875
precision	0.9063	0.871	0.8824	0.9032	0.871

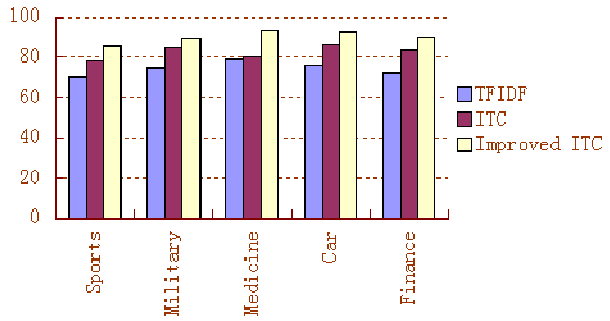

Figure 1. F-scores of the three Feature Selection Algorithms.

As shown from the result shown by **Tables 1** and **2**. In the aspect of short text classification, the ITC algorithm may improve the performance of the TFIDF up to because of the term frequency weakness of the ITC. Thus as shown from **Tables 2** and **3**, the Improved ITC raises the precision rate and recall rate obviously, so the Improved ITC boost the performance a certain extent. In addition, there are many aspects require further investigations though the short text classification, like the fact that we conducted that the predefined category is more professional and the classification result is more satisfied. For instance, the category medicine is more professional relative to the other categories, so the performance of the medicine category is pretty good even we only adopt the conventional TFIDF algorithm. If we do not choose the finance category, and we choose the stock subcategory of the finance category, we will find that the classification result of this category is more accurate than the finance category, so the contradiction problem of the profession and universality of predefined category is needed to be discussed in the future work, this situation provides new improvement space for the short text classification. In conclusion, we specially have considered the structural feature of the short documents, then adopt the ITC feature selection algorithm, and improve the ITC on the basis. The experimental results proved that our improvements are very successful for the short text classification. Moreover, the results of the new feature selection algorithm also confirm our guess about the ITC which rare researches adopt that the term frequency effect of the short documents which appearing in have to be weakened.

5. Conclusion and Future Work

In this paper, we mainly concentrate on the study of the short text classification, which is developing in the field of text classification. Firstly, we explore the structural and language characteristics which the short documents own particularly, then we decide to give up the conventional TFIDF but adopt the ITC feature selection algorithm based on the language characteristics of the short documents. The effect of term frequency is weakened compared with the long documents. We propose two new functions creatively based on the concept of information entropy and the weight effect of the term position factor. Lastly we improve the conventional ITC on the basis of the concluded defects of the ITC algorithm and the structural characteristics of the short documents with the two functions: Documents Distribution Entropy and Position Distribution Weight. The experiment results show that the measures we proposed are helpful for the short text feature selection, and the proposed feature selection algorithm in this paper can play an effective role in the field of the short text classification. Moreover, there are additional aspects that deserve further researches. In the previous experiments, how to solve the contradiction problem of the profession and universality when we predefine the training documents categories? Then, we will try to adjust the position weight value and hope to confirm the best value, and we also consider if there are other ways to combine the two new functions with the ITC. The future work will make the performance of the new algorithm boost more.

6. Acknowledgements

The authors are grateful to the School of Information Science and Engineering, University of Jinan which provides a huge amount of data used in this study. To Qu Shouning and Du Tao (School of Information Science and Engineering, University of Jinan) who provided many useful suggestions.

REFERENCES

[1] Z. Y. Cui, “Study on Related Technologies of Chinese

Short Text Classification," Henan University, Henan, 2006.

[2] D. Fan, "Study on Chinese Short-Text Classification," Tsinghua University, Beijing, 2009.

[3] G. Salton, "Antomatic Text Processing: The Information, Analysis, and Retrieval of Information by Computer," Addison-Wesley Longman Publishing Co., Inc., Boston, 1989.

[4] V. Hatzivassiloglou, J. Klavans and E. Eskin, "Detecting Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine," *Proceedings of Joint SIGDAT Conference on Empirical Methods in NLP and very Large Corpora*, Hong Kong, June 1999, pp. 21-22.

[5] J. Chang, "Study on Short Text Classification Algorithms," Fudan University, Shanghai, 2008.

[6] W. Y. Liu, J. Q. Xiao, F. Min and B. Liu, "A Short Text Modeling Method Combing Semantic and Statistic Information," *Information Science*, Vol. 180, No. 20, 2010, pp. 4031-4041. http://dx.doi.org/10.1016/j.ins.2010.06.021

[7] X. Q. Wu, "Application of Hierarchical Keyword Extracting and Text Classification in BBS," Shanghai Jiao Tong University, Shanghai, 2006.

[8] S. Wang, X. H. Fan and X. L. Chen, "Chinese Short Text Classification Based on Hyponymy Relations," *Journal of Computer Application*, Vol. 30, No. 3, 2010, pp. 602-606.

[9] Z. Y. Cui, "Microblog Text Classification Based on Semantic Information," *Modern Computer*, Vol. 8, 2010, pp. 18-20.

[10] Z. M. Han, Y. S. Zhang, H. Zhang, Y. L. Wang and J. H. Huang, "On Offensive Short Text Tendency Classification Algorithm for Chinese Microblog," *Computer Application and Software*, Vol. 29, No. 10, 2010, pp. 89-103.

[11] Z. F. Zhang, D. Q. Miao and G. Gao, "Short Text Classification Based on LDA Topic Model," Computer Application, Vol. 6, 2013, pp. 1587-1590

[12] G. L. Shi and Q. F. Shi, "Text Mining Based on Consistency of Product Reviews in Different Shopping Websites," *New Technology of Library and Information Service*, Vol. 12, 2011, pp. 64-68.

[13] W. Yi and C. Meek, "Improving Similarity Measures for Short Segments of Text," *Proceedings of* 22*nd conference on Artifical Intelligince* (*AAAI*-07), Vancouver, 24-26 July 2007, pp. 1489-1494.

[14] Y. T. Zhou, J. B. Tang and J. Q. Wang, "The Improved TFIDF Based on Information Entropy," *Computer Engineering and Application*, Vol. 43, No. 35, 2007, pp. 156-158.

[15] K. L. Chen, "Collection and Analysis of Large-Scale Balanced-Corpus and Approach to Text Categorization," Chinese Academy Science, Beijing, 2006.

[16] Y. F. Zhang, S. M. Peng and J. Lv, "The Improvement and Application of TFIDF Based on Text Classification," *Computer Engineering*, Vol. 3, No. 19, pp. 76-78.

[17] F. J. Shao and Z. Q. Yu, "Principle and Algorithm of Data Mining," China Waterpower Press, Beijing, 2003.

[18] J. Qin, X. R. Chen and W. J. Wang, "Feature Extraction of Text Classification," *Computer Application*, Vol. 23, No. 2, 2003, pp. 45-46.

[19] W. H. Jia and M. Kamler, "Data Mining: Concepts and Techniques," Morgan Kaufman Publishers, New York, 2006.

