Int. J. Chem. Sci.: 14(1), 2016, 449-457
ISSN 0972-768X
www.sadgurupublications.com

A MODIFIED FREQUENCY BASED TERM WEIGHTING APPROACH FOR INFORMATION RETRIEVAL

M. SANTHANAKUMAR[a*] and C. CHRISTOPHER COLUMBUS[b]

Department of Computer Science and Engineering, PSN College of Engineering and Technology, Melathediyoor, TIRUNELVELI – 627152 (T.N.) INDIA

ABSTRACT

Term frequency-inverse document frequency (TF-IDF) is one of the repeatedly used term weighting methods, which assigns weights based on the occurrences of a term in a document. This paper proposes an improved TF-IDF method using multi term occurrences in a document. To achieve the best performance, pre-processing methods such as tokenization, stopword removal and stemming are applied on both user query and document terms. The experimental results of the proposed work are compared with existing term weighting methods such as TF, IDF, TF-IDF and entropy. The proposed method gives better average precision, recall and F-score values than the existing methods.

Key words: TF, IDF, Entropy, WWW, Term weighting.

INTRODUCTION

Lately, due to the amount of information has been rapidly growing on the web, it is essential that the user can access information without losing any data related to them. At the same time, problems like unnecessary details may occur while retrieving a document from high quantity of data. A fair quantity of effort has been done in Data Mining (DM) for retrieving relevant documents that the user requires. Vector Space Model (VSM) is a frequently used model based on the use of index terms. The index term denotes the importance of a term by assigning weight to that term. It is called as term weighting. TF-IDF is a frequently used method for assigning weights to the term based on its occurrences in a document[1]. Term Frequency (TF) is determined by the frequency of occurrence of a term in a document or collection of documents. Inverse Document Frequency (IDF) is another measure to determine the number of documents, which contains the term[2]. In recent times, there are many modified TF-IDF algorithms proposed and implemented[3]. Before assigning

Author for correspondence; E-mail: santhanakumar@psncet.ac.in, columbus@psncet.ac.in

weight, pre-processing techniques such as tokenization[4], stopword removal[5] and stemming[6] are applied to remove the unnecessary data from the documents. This paper, proposes a modified TF-IDF term weighting scheme based on classical TF-IDF and the performance of the proposed work is measured by average precision, average recall and average F-score values.

Related works

Liu et al.[7], proposed an improved term weighting method named Term frequency Inverse Positive Negative Document frequency (TFIPNDF), which exposes the importance of a term based on the distribution in positive-negative training sets. Lan et al.[8], proposed a new method called Term Frequency-Relative Frequency (TF-RF), which is based on a term frequently distributed in a document. Xia and Chai[9], developed a method to assign weights by considering local and global term weighting scheme. The higher weight is given to the term that is uniformly distributed and widely appeared inside the document. Gautham and Kumar[10], implemented a new method assigning higher weight to the rare term even it has low frequency based on class information such as intra class and inner class information. Wang and Zhang[11], proposed two different novel term weighting methods named Inverse Category Frequency (ICF) and TF-ICF based on traditional TF-IDF. Goswami et al.[12], developed a new method named Document Frequency-Inverse Corpus Frequency (DF-ICF). In DF, the frequently viewed document by the user gets higher frequency. ICF calculates the number of corpus containing the document. Sabbah et al.[13], implemented a modified TF-IDF method and compared with existing term weighting scheme using the dataset dark web content. This method gave better precision, recall, and accuracy values than the existing methods.

From the above study it is clearly identified number of modified term weighting methods were proposed and implemented to overcome the drawbacks of classical TF-IDF method to predict the user related information. Likewise this paper proposes a new modified frequency based term weighting method to achieve the better performance when compared with classical TF-IDF.

EXPERIMENTAL

Dataset

In order to check the effectiveness and retrieval accuracy of the proposed work 20 Newsgroup dataset collections[14] are used. It contains nearly 20,000 documents in it. Each category contains around 1000 documents. To assess the performance of proposed method

10 categories from 20 Newsgroup dataset have been selected and split into test and training sets. Sample queries were generated randomly using test set and experiments have been done using training set. Table 1 describes the selected 10 categories and shows the number of documents in each category. The documents in each category are pre-processed for tokenizing, stopword removal and stemming. The multiple Query Vector (QV) is prepared randomly from the test dataset. The QV is described in equation 1.

$$QV = (qt_1, qt_2 \ldots\ldots qt_n) \quad \ldots(1)$$

Here n is the number of unique terms and t_i (i = 1, 2,…n) denotes each term of QV.

Table 1: Number of documents in each topic from 20 Newsgroup training set

Category	**alt. atheism**	**Comp. graphics**	**Comp.os.ms-windows.misc**	**Misc. forsale**	**Rec. motorcycles**
No. of Documents	480	584	587	585	598
Category	Rec.sport. hockey	Sci. electronics	Sci. space	Soc.religion. christian	Talk.politics. mideast
No. of Documents	600	591	593	599	397

Proposed term weighting method

TF-IDF is the most frequent term weighting method for assigning weight to the terms in the field of IR, classification and clustering. This proposed work is an extended version of classical TF formula and it is denoted by Imp-$TF_{t,d}$. The proposed formula is expressed in Equation 2.

$$\text{Imp-}TF_{t,d} = \frac{TF_{t,d} \times ATF_{t,d}}{M} \quad \ldots(2)$$

Here, the term $ATF_{t,d}$ describes the average frequency of the term t_i in a corpus. Denominator M is expressed in equation 3.

$$M = \frac{L(d_j)}{D(d_k)} \quad \ldots(3)$$

Where, L denotes the document length of d_j and D describes the total number of distinct terms in a corpus d_k.

As in classical TF-IDF, $tf_{t,d}$ denotes the number of occurrences of the term t_i in the document d_j. However, $ATF_{t,d}$ describes the average term frequency of the term t_i expressed in equation 4.

$$ATF_{t,d} = \frac{\sum_{d=1}^{k} tf_{t,d}}{N} \quad \ldots(4)$$

Where, k and N denote the total number of documents in a corpus and total number of documents that containing the term t_i respectively.

In this proposed work, the classical TF formula is replaced with the improved TF formula in TF-IDF term weighting method, and which is expressed in equation 5 and 6.

$$\text{Imp TF-IDF}_{t,d} = \text{Imp TF}_{t,d} \times IDF_t \quad \ldots(5)$$

$$\text{Imp TF-IDFt,d} = \left(\frac{TF_{t,d} \times ATF_{t,d}}{\left(\frac{L(d_j)}{D(d_k)} \right)} \right) \times IDF_t \quad \ldots(6)$$

RESULTS AND DISCUSSION

To evaluate the efficiency of the proposed work, different QV's are formulated randomly from the test set according to the category. The outcomes of improved TF-IDF are compared with other term weighting methods such as TF, IDF, TF-IDF and Entropy. The average precision, recall and F-score values of different term weighting method are described in Table 2 and it clearly shows that the improved TF-IDF has a better precision, recall and F-score values than other term weighting methods.

The performance measures of different term weighting methods are described in Fig. 1, Fig. 2 and Fig. 3, respectively. Fig. 4 shows the average precision recall values of improved TF-IDF method. From Fig. 1 it can be clearly identified in most of the categories the proposed term weighting method has higher average precision values than other term weighting methods, which are available in the literature.

Table 2: Performance measures of proposed and existing term weighting methods

		Alt. atheism	Comp. graphics	Rec. motor-cycles	Sci. electronics	Misc. forsale	Talk. politics. mideast	Soc. religion. christian	Comp. Os.ms - windows. misc	Rec. sport. hockey	Sci. space
TF	Precision	0.9523	0.8293	0.9756	0.9106	0.8611	0.7313	0.7105	0.9985	0.9956	0.9421
	Recall	0.7150	0.7689	0.9302	0.8498	0.7356	0.6561	0.9153	0.5833	0.9286	0.8905
	F-score	0.8168	0.7980	0.9524	0.8792	0.7934	0.6917	0.8000	0.7364	0.9609	0.9156
IDF	Precision	0.9653	0.8333	0.9663	0.7282	0.8767	0.8420	0.7485	0.9892	0.9948	0.9462
	Recall	1.0000	0.5111	1.0000	1.000	0.6508	0.6087	0.4689	0.5278	0.8980	0.4378
	F-score	0.9823	0.6336	0.9829	0.8427	0.7470	0.7066	0.5766	0.6883	0.9439	0.5986
TF-IDF	Precision	0.9679	0.7441	0.9652	0.7195	0.9663	0.9727	0.9597	0.9166	0.9851	0.9592
	Recall	0.9860	0.9883	1.0000	1.0000	1.0000	0.9817	0.9917	0.9930	0.9167	0.9826
	F-score	0.9769	0.8490	0.9823	0.8369	0.9829	0.9772	0.9754	0.9533	0.9497	0.9708
Entropy	Precision	0.9683	0.7377	0.9773	0.8905	0.8912	0.8650	0.7254	0.9863	0.9703	0.9263
	Recall	0.9924	0.9891	0.9783	0.7390	1.0000	0.6087	0.9467	0.9926	1.0000	0.9536
	F-score	0.9802	0.8451	0.9778	0.8077	0.9425	0.7146	0.8214	0.9894	0.9849	0.9398
Proposed	Precision	0.9683	0.8520	0.9785	0.9224	0.9763	0.9732	0.9662	0.9268	0.9963	0.9625
	Recall	1.0000	1.0000	0.9987	1.0000	0.9989	0.9983	1.0000	1.0000	1.0000	0.9998
	F-score	0.9839	0.9201	0.9885	0.9596	0.9875	0.9856	0.9828	0.9620	0.9981	0.9808

The average precision values of categories, rec.motorcycles and rec.sport.hockey are close to the proposed term weighting method. However, the proposed method has higher average precision value than the other methods for all other categories. From this it is clearly understood most number of relevant documents are retrieved from the number of retrieved document based on multiple queries.

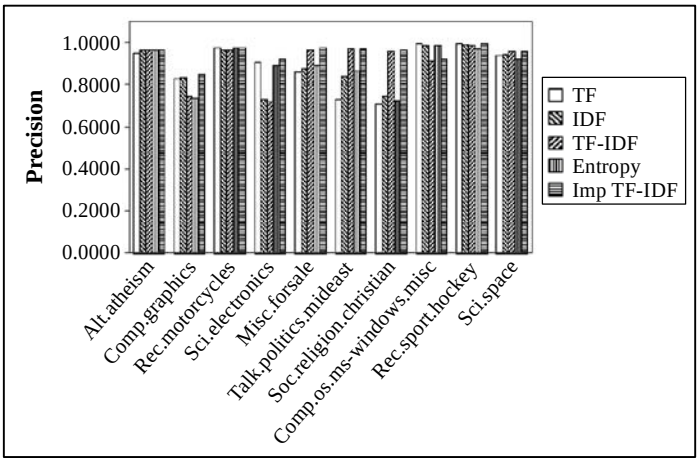

Fig. 1: Average precision of improved TF-IDF and other term weighting methods

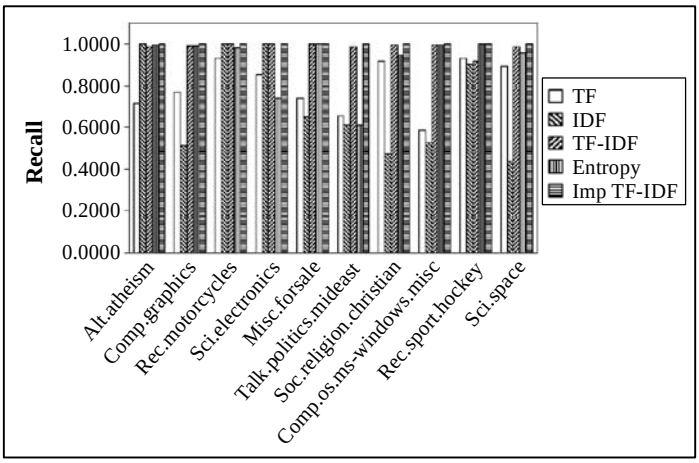

Fig. 2: Average recall of improved TF-IDF and other term weighting methods

Fig. 2 shows the performance evaluation of average recall values for all term weighting methods. When comparing to existing methods, the proposed term weighting method has highest value in most of the categories. For the categories soc. religion. christian and comp. graphics the average recall value of proposed method is approximately double

than that IDF. In the same way, in most of the categories the method entropy has the closest value with proposed method. However, the proposed term weighting method has the higher average recall value in all categories except the category misc. forsale. The completeness of the proposed method is ensured by the highest recall values.

Another measure called F-Score has been used to assess the effectiveness of the proposed scheme. F-score is the combination of precision and recall measurement. The average F-score values of all term weighting schemes are shown in Fig. 3. In categories alt.atheism, rec.motorcycles, talk. politics. mideast, misc. forsale, soc.religion. christian and sci.space the average F-Score value of TF-IDF is close to the proposed term weighting method.

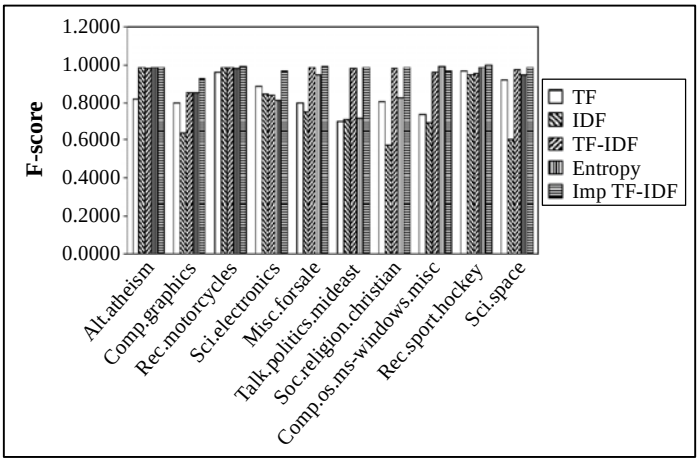

Fig. 3: Average F-score of improved TF-IDF and other term weighting methods

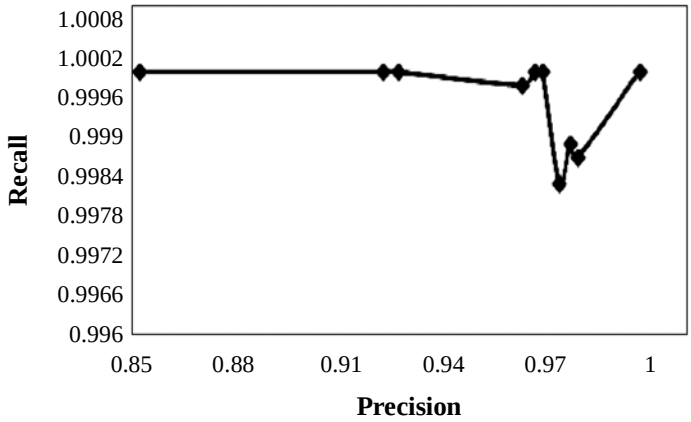

Fig. 4: Average precision/Recall graph for improved TF-IDF

Even though, TF-IDF has lesser average F-score value than the proposed method. Likewise, in categories soc.religion.christian and sci.space the average F-score value is almost doubled than IDF method. Fig. 4, it is a plot of average precision Vs recall value and also it shows the best retrieval accuracy for proposed term weighting method. From these, it can be inferred that the proposed term weighting method clearly predicts more accurate feature vectors of web pages than other term weighting methods, which are discussed in section 3.

CONCLUSION

This paper proposed an improved term weighting method to assign weights to the terms and retrieve the relevant document based on user query. It is the extended work of classical TF-IDF method. In this proposed work, average frequency of a term is calculated and multiplied with classical TF formula. For normalizing the length, the proportion of each document length and total number of unique words are calculated. To prove the better performance of proposed method, 20 Newsgroup dataset is used. From the testing set user queries are randomly generated and experiments are done with the training set. The outcome of the proposed method has better precision, recall and F-score values than the existing methods. In future, by applying this proposed method conduct more experiments in different datasets to validate the different classifiers and clustering methods.

ACKNOWLEDGEMENT

The first author thanks to University Grants Commission (UGC), India for providing financial support under Rajiv Gandhi National Fellowship from 2013-15.

REFERENCES

1. D. L. Lee, H. Chuang and K. Seamons, Document Ranking and the Vector-Space Model, IEEE Software, **XIV(2)**, 67-75 (1997).
2. G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company (1988).
3. M. Santhanakumar and C. C. Columbus, Various Improved TFIDF Schemes for Term Weighting in Text Categorization: A Survey, Int. J. Appl. Engg. Res., **X(14)**, 11905-11910 (2015).
4. W. R. W. Zulkifeli, N. Mustapha and A. Mustapha, Classic Term Weighting Technique for Mining Web Content Outliers, International Conference on Computational Techniques and Artificial Intelligence, Penang, Malaysia, 271-275 (2012).

5. X. Zhu, Basic text Process, Advanced Natural Language Processing, Spring (2010). [Online] Availablehttp://pages.cs.wisc.edu/~jerryzhu/cs769/text_preprocessing.pdf (current July 2015).

6. M. F. Porter, An Algorithm for Suffix Stripping, Program, **XIV(3)**, 130-137 (1980).

7. L. Liu and T. Peng, Clustering-Based Method for Positive and Unlabeled Text Categorization Enhanced by Improved TFIDF, J. Information Sci. Engg., **XXX(5)**, 1463-1481 (2014).

8. M. Lan, C. L. Tan, J. Su and Y. Lu, Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, Pattern Analysis and Machine Intelligence, **XXXI(4)**, 721-735 (2008).

9. T. Xia and Y. Chai, An Improvement to TF-IDF: Term Distribution Based Term Weight Algorithm, J. Software, **VI(3)**, 413-420 (2011).

10. J. Gautham and E. Kumar, An Integrated and Improved Approach to Terms Weighting in Text Classification, Int. J. Computer Sci., **X(1)**, 310-314 (2013).

11. D. Wang and H. Zhang, Inverse-Category-Frequency Based Supervised Term Weighting Schemes for Text Categorization, J. Information Sci. Engg., XXIX, 209-225 (2013).

12. P. Goswami and V. Kamath, The DF-IF Algorithm-Modified TF-IDF, Int. J. Computer Applications, **XCIII(13)**, 28-30 (2014).

13. T. Sabbah and A. Selamat, Modified Frequency-Based Term Weighting Scheme for Accurate Dark Web Content Classification, In proceedings of the 10th Asia Information Retrieval Societies Conference, Springer, 184-196 (2014).

14. K. Lang, Newsweeder: Learning of Filter Netnews, In Proceedings of the Twelfth International Conference on Machine Learning, 331-339 (1995).

Accepted : 16.03.2016