Term Weighting Using Term Dependence

Raj Kishor Bisht
Dept. of Computer Science
Amrapali Institute of Management and Computer Application,
Haldwani (Uttarakhand)-India

Garima Srivastava
Dept. of Computer Science
Amrapali Institute of Management and Computer Application,
Haldwani (Uttarakhand)-India

H. S. Dhami
Dept. of Mathematics
University of Kumaun,
S.S.J.Campus Almora
(Uttarakhand) -India

ABSTRACT
Performance of an information retrieval system depends on its weighting scheme. Weighting of a term can be seen in two aspects, local and global. For each type of weighting scheme, generally, single terms are considered. Term dependency is quite natural in a document. Word pairs or phrases can better describe a document in place of single terms. In the present paper an attempt has been made to study and quantify the dependency of terms to each other. Term dependency has been utilized to define a local weighting scheme for word pairs. Utility of the proposed weighting scheme has been shown by arbitrarily choosing some documents and extracting relevant word pairs from the documents.

General Terms
Information Retrieval, Data Mining

Keywords
Weighting scheme, Local weight, Global weight

1. INTRODUCTION
Term weighting is an integral part of Information retrieval system which plays an important role in the performance of the information retrieval system. Weight of a term provides the information about the relevance of the term to the document. Weight of a term can be seen in two different aspects, local and global. Local weights are functions of frequency of a term in a document and global weights are functions of the frequency of a term in the entire collection. Manning & Schutze [6] and Sandor Dominich [3] have provided a detail of information retrieval techniques and weighting schemes in their books. A detailed discussion on different types of local and global weighting schemes can be found in Chisholm & Kolda [2]. Suitability of some global weighting schemes for local weights can be seen in Bisht & Dhami [1]. A comparative study on term weighting schemes for text categorization has been made by Lan et al [5]. There are several weighting schemes available in the literature. Following are some of the commonly used local weighting schemes:

Binary: The simplest local weighting scheme is the Binary weighting scheme which can be expressed as follows:

$$w_{ij} = \begin{cases} 1 & if \quad f_{ij} > 0 \\ 0 & if \quad f_{ij} = 0 \end{cases} \tag{1}$$

where f_{ij} is the frequency of term i in document j.

Frequency weighting scheme: It is the frequency of a term in a given document.

$$w_{ij} = f_{ij} \tag{2}$$

Logarithm of frequency: Term frequency with in a document is not a relatively good descriptor of a term. For example, if a term i appears 15 times in a document and term j appears single time, then according to their frequency weights, the i^{th} term should be 15 time important than j^{th} term but it is not necessary. When we take logarithms of frequency, it becomes relatively good descriptor of the importance of terms. It is defined as follows:

$$w_{ij} = \begin{cases} \log(1 + f_{ij}) & if \quad f_{ij} > 0 \\ 0 & if \quad f_{ij} = 0 \end{cases} \tag{3}$$

Normalization of weighting schemes is also used to get the weight of terms in a certain interval, that is [0, 1]. In this type of weighting scheme, we divide the weighting scheme by a suitable normalization factor. As an example, Maximum f_{ij}, where $1 \le i \le n$ may be a normalization factor for frequency weighting scheme if there are total of n terms in the document.

Term weighting schemes are generally defined on the basic assumption that a term is independent to other terms. Term weighting schemes based on the assumption of term to term independence may have a chance of information of a document being lost. Kim Hee-Soo et al [4] computed term dependencies to refine global term weighting. Term dependency in a document is also quite natural. In place of a single term a word pair may be a better key word of a document and also can better describe a document.

In the present paper, we have proposed a local weighting scheme for word pairs which exhibits that a word pair may be a better key word rather than a single word.

2. PROPOSED WEIGHTING SCHEME FOR WORD PAIRS
We have proposed a local weighting scheme which is based on the individual weight of words in the pair and the closeness between the two words. The weighting scheme for pair of words can be understood through the following steps:

1. Select a document for weighting scheme.

2. Remove all function words, that is, articles, determiners, propositions, helping verbs etc.
3. Transform word to their lexical roots. For example the words 'prisoner', 'prisoners' have been transformed to 'prisoner'.
4. Count the frequency of each non function word in the document. For a word w_1 the frequency shall be counted as follows:
   $f(w_1)$ = No. of sentences in which w_1 appears.
5. Select some (five to ten) most frequent word from the list.
6. Make the combination of most frequent words and count the frequency of word pairs. Frequency of each combination shall be counted as follows:
   $f(w_1, w_2)$ = No. of sentences in which both the words w_1 and w_2 appear.
7. For each word pair (w_1, w_2), calculate the probability that the word combination is relevant to the document, denoted by $P_R(w_1, w_2)$ and defined as follows:

$$P_R(w_1, w_2) = \frac{f(w_1, w_2)}{Min[f(w_1), f(w_2)]} \quad (4)$$

Since $f(w_1,w_2)$ can not be greater than $Min\ [f(w_1), f(w_2)]$, therefore, $Min[f(w_1), f(w_2)]$, represents maximum possible occurrence of (w_1,w_2) while $f(w_1,w_2)$ shows the actual occurrence of (w_1,w_2). Thus, $P_R(w_1,w_2)$ shows the probability that the words w_1 and w_2 being appear together and hence, probability that the word pair is relevant to the document.

8. Calculate weight of word pair (w_1, w_2) defined as follows:

$$W(w_1, w_2) = [\log(1 + f(w_1)) + \log(1 + f(w_2))] * P_R(w_1, w_2) \quad (5)$$

The proposed weighting scheme is a combination of logarithm of frequency of each word multiplied by the weight of the word pair in the form of probability of the word pair being relevant to the document.

3. EXPERIMENTAL RESULTS

For the purpose of experiment, we have arbitrarily chosen three different texts; first one "Gandhi the Prisoner", second "Gandhi: Soldier of Nonviolence" and third "The Spiritual Basis of Satyagraha" (see appendix). Weights of some of the frequent word pairs have been calculated. Table 1 exhibits the weights of word pairs for different documents.

Table 1. Weight of word pairs in the documents

Document I

w_1	w_2	$f(w_1, w_2)$	$W(w_1, w_2)$
Political	Prison	6	2.03
Gandhi	Imprisonment	4	1.20
Gandhi	Prison	13	0.92
Cell	Prison	3	0.77
Gandhi	cell	3	0.76
Gandhi	Indian	5	0.72
Prison	Labour	2	0.76

Document II

w_1	w_2	$f(w_1, w_2)$	$W(w_1, w_2)$
Gandhi	India	11	2.67
Gandhi	Africa	8	2.30
Gandhi	Salt	4	1.62
Gandhi	Protest	5	1.44
Gandhi	March	3	1.42
Gandhi	Satyagraha	3	1.42
Salt	March	3	0.97

Document III

w_1	w_2	$f(w_1, w_2)$	$W(w_1, w_2)$
Truth	God	4	0.77
Satyagraha	Non -violence	2	0.53
Satyagraha	Gandhi	2	0.53
Satyagraha	Truth	2	0.34
Gandhi	Truth	2	0.34
Satyagraha	People	1	0.21
Satyagraha	God	1	0.21

The first document is mainly about Mahatma Gandhi and his prison conditions, second document contains general information about Gandhiji and the third one is concerned mainly with Gandhiji's way of Satyagraha and spiritualism. In the first document some highest frequency words are Prison-63, Gandhi-49, Indian-21, cell-11, warder-10, Imprisonment-9, opinion-8, political-8, laws-7 and labour-7. In the second document these are Gandhi-38, India-11, Africa-9, South-9, Protest- 9, Salt-6, Cotton-6, Law-6, March-5, Make-5, Satyagraha-5, Independent-5 and in the third document frequent words are Satyagraha-23, Truth-15, People-12, God-12, Patient-10, Rogers-10, Unity-10,Non-violence-9, Violent-9, Gandhi-9. From the table 1, it can be observed that the words 'Political' and 'Imprisonment' have low frequencies yet they make an important combination with other words. In the third document the word pair (Truth, God) has highest weight which itself proves the content of the document. Weights of word pair also show the dependency of words in the pair on each other.

4. CONCLUSION

In the present paper, it has been shown that word pair can better describe a document. Though single terms may describe the

content of a document, yet a word pair with its relevance to a document provides a clear idea about the content. The first two documents are related to Gandhi and frequency weights of the two words in the documents are almost similar. In this case, combination of the word 'Gandhi' with other words may play an important role. Table 1 shows that the relevant word pairs have higher weight than other word pairs and have the capacity to describe the document. Hence, the proposed weighting scheme for word pair is quite useful. The proposed weighting can also be extended to a combination of more than two words or a phrase.

5. REFERENCES

[1] Bisht R.K. and Dhami H S. 2008. On some properties of content words in a document. In Proceedings of the 6th Annual conference of Information Science and Technology Management, 51 (1-19).

[2] Chisholm, E. and Kolda T. G. 1999. New term weighting formulas for the vector space method in information retrieval. Technical report ORNL-TM-13756, Oak Ridge national laboratory, Oak ridge, TN.

[3] Dominich S. 2008. The Modern Algebra of Information Retrieval (Information Retrieval Series). Springer-verlag, New York.

[4] Kim H., Choi I. and Kim M. 2004. Refining Term Weights of Documents Using Term Dependencies. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 552-553.

[5] Lan M., Sung S. Y., Low H. B. and Tan C. L. 2005. A comparative study on term weighting schemes for text categorization. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vol 1, 546-551.

[6] Manning C. D. and Schutze H. 2002. Foundations of Statistical Natural Language Processing. MIT press, Cambridge.