The Effect of Term Importance Degree on Text Retrieval

Soheila Karbasi
Department of Computer Science, Golestan University, 49138-15759 Gorgan, Iran

Mehdi Yaghoubi
Department of Computer Science, Golestan University, 49138-15759 Gorgan, Iran

ABSTRACT

Various approaches to index term-weighting have been investigated. In fact, term-weighting is an indispensable process for document ranking in most retrieval systems. As well actual information retrieval systems have to deal with explosive growth of documents of various sizes and terms of various frequencies because an appropriate term-weighting scheme has a crucial impact on the overall performance of systems.

This paper attempts to investigate the impact of term-weighting parameters used in the most well-known retrieval models. The study has been particularly focused on normalization of term frequency in weighting schemes. A novel factor which is called "term importance degree" has been identified, which can be applied to term-weighting schemes by using several parameters. The calculated correlations between the parameters of weighting schemes confirmed the impact of this factor to increase the performance of text retrieval systems. Two models of term frequency normalization are inserted in a basic term-weighting scheme, which shows the importance of terms. The experiments were carried out on the standard test collections which validated by multiple statistical tests.

General Terms

Information system, Indexing, Information retrieval

Keywords

Text retrieval, Term-weighting scheme, Term frequency normalization, Term importance degree

1. INTRODUCTION

Many studies show that most information is text based and high retrieval performance is closely related to the term-weighting schemes [3]. Hence, continuously increasing the number of text documents, intranets and digital libraries leads to the use of more efficient and effective retrieval methods.

Many term-weighting approaches have been proposed and studied in the domain of textual information retrieval so that term-weighting process should provide an instructive representation of content of documents. Also, it should supply an indicator of importance to discriminate the terms (indexing units). Term-weighting has been explained by controlling the exhaustivity and specificity of the search, where the exhaustivity is related to recall and specificity is related to precision [14].

In this paper, BM25 weighting scheme and its parameters efficiency were investigated to propose new parameters supporting the tf-idf measures. Particularly, it focused on redefining the representation of documents and mostly focused on the weights that assigned to indexed terms. It presents the importance of terms in documents based on term frequency, and a specific normalization factor of each document which is independent from other documents in the collection. To offer new scheme, an empirical study on two TREC collections containing the Web pages is conducted. The performance of BM25 model with various combinations of its parameters is analyzed in order to calculate the degree of their influence on the retrieval performance. The first goal of this experiment was to assess the impact of term-weighting parameters based on performance (recall/precision). Using statistical techniques, two points have been specially attended. The first point was to measure the correlations coefficients between the parameters and the second point was to evaluate the distribution of document lengths in the collections. It was realised that the correlations between all parameters used in weighting model were weak. Also, a particular term with a high frequency is not necessarily in a long document which means the term frequency will be penalized by classic methods of document length normalization.

Based on preliminary results, a new factor called "term importance degree" is introduced. This factor takes into account the location of terms in the ranked list by decreasing their frequencies in documents. So, a term can have the same degree of importance in two documents even though their frequencies aren't the same. It is proposed to use this factor to adjust both the effect of term frequency and document length and introduced two new parameters which are used in new term-weighting schemes. The first was called "Balanced term importance degree" which revises the interaction between frequency and term importance degrees. This parameter improves the degree of top terms of each document, and consequently a frequent but not important term (not presents in the top terms) will not be boosted. The second parameter "Average term importance degree" simply replaces document length parameter. This parameter represents the average of term importance degrees of documents. It will be unique in each document and independent from the collection and the number of documents in collections. In order to evaluate the performance of presented schemes, they have tested with test collections and compared with baseline results. The results clearly showed that their performances are comparable with BM25. The rest of the paper is organized as follows: Section 2 reviews the popular term-weighting models. Section 3, proposes the experimental settings. Experiments and results are described in Section 4 and Section 5 presents new parameters with relation experiments. Section 6 contains the discussion and conclusions.

2. TERM-WEIGHTING SCHEMES

Several models are proposed in the literature for term-weighting which is the core of any information retrieval system. They determine the important terms of documents to estimate the relevance of a document to a query. In spite of the recent progress in information retrieval techniques, the performance of text based retrieval systems is largely dependent on term-weighting models. In addition, large-scale retrieval performance requires the use of appropriate term-weighting scheme since it dominates the computational demands of retrieval [3]. In general, term-weighting schemes are based on statistical [13, 9], semantic [15] or probabilistic models [10]. The most well-known are Okapi [14], Lnu [4], dtu [19], Pivoted normalization [17], Simplified Similarity Scoring [2], PL2 [1], ETW [7] that are used by the various IR systems.

One of the most commonly used term-weighting schemes is tf-idf model that is based on two basic principles of term-weighting:

– For a given term in a document, the higher term frequency is, the more likely the term is relevant to the document,

– For a given term, the higher the term occurs throughout all documents, the less likely the term discriminates between documents [16].

There are numerous variants of tf-idf weighting scheme, while most can be described as particular cases of the initial introduced by Salton and Buckley in the SMART project [11, 19, 16]. Most of them are parametric and a fixed form of density function with parameters that dependent upon the number of documents and their sizes [8]. Each scheme can be represented as a triple of parameters XYZ, where X stands for the term frequency parameter, Y for the document frequency, and Z for the normalization parameter.

The evolution of 2-Poisson model as designed by Robertson, Van Rijsbergen and Porter has motivated the birth of BMs family term-weighting scheme (BM for Best Match). BM25 is one of the most well established tf-idf weighting models [12, 13] which is introduced as follow:

$$W = \sum_{t \in q \cap d} \frac{tf}{tf + k_1 n_b} log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) qtf \quad (1)$$

Where:

tf: frequency of term t in the document d
qtf: frequency of term t in the query
df$_t$: number of documents containing term t
N: total number of documents in the collection
k$_1$: controlling parameter that is set with 1.2
n$_b$: normalization factor that is calculated as:

$$n_b = (1 - b) + b\frac{dl}{Avgdl} \qquad (2)$$

dl: length of document d, in tokens
Avgdl: average documents length, in tokens
b: tuning parameter ($0 \le b \le 1$)

3. EXPERIMENTAL SETTING

The key point of this study was to evaluate the performance of parameters used in BM25 scheme (tf, df and length of document), in relation with collection size. In the following, the study of the problems caused by these parameters in IR and some suggested approaches for dealing with them are presented.

For evaluation, we used Wt2g and Wt10g collections whose details are shown in Table1. Both collections consist of documents from the Web which are distributed by CSIRO[1]. The documents of collections are presented in SGML or HTML format [6]. The queries are also issues of Topics provided by TREC whose Title and Description fields are considered. The statistics on query sets of each collection are shown in Table 2.

Table 1. Collections characteristics

Collection	*Wt2g*	*Wt10g*
# Documents	248,440	1,677,562
# Indexed terms	1,212,289	3,095,678
Avg. # terms per document	259	282
Avg. # unique terms per document	133	151
Collection size (GB)	2	10

Table 2. Topics statistics

Collection	*Topic set*	*Avg. # term per query*
Wt2g	401 - 450	*6*
Wt10g	501 - 550	*4*

4. INITIAL EXPERIMENTS

4.1 Term-Weighting Parameters Impact

To compare the effect of term-weighing parameters, the performance of BM25 with various combinations of its parameters is evaluated. Table 3 shows the MAPs [5] obtained with combinations of various parameters based on the following formulas. The results are performed with no query expansion and BM25 model is used with k_1=2 and b=0.75 in all experiments.

$$W = \frac{tf \times \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right)}{2 + tf} \qquad (3)$$

$$W = \frac{tf}{2 \times \left(0.25 + 0.75\frac{dl}{Avgdl}\right) + tf} \qquad \mathit{(4)}$$

Table 3. Values of MAP with various parameters combination of BM25

Collection	*formula 3*	*formula 4*	*BM25 formula*
Wt2g	0.0497	0.2307	0.2635
Wt10g	0.0379	0.1676	0.1884

According to the results, it may be supposed that normalized document length impact is more important than df parameter. In fact, the importance of document length normalization is a significant topic in term-weighing and most well-known normalization methods, such as maximum term frequency, pivoted normalization, and byte length normalization use document length normalization for term frequency parameter normalization [18].

Based on this analysis, it is tried to nominate an auxiliary factor for document length normalization in order to better documents discrimination. Therefore, the relationship (dependency) between term frequency and the other parameters of weighting scheme was analyzed. The relationships are measured based on the variance and covariance values of statistical methods. 50 indexed terms of test collections have been selected randomly. The correlation between their frequency and length of documents in related documents are calculated. The results show that there is not high correlation between tf and document length parameters. The maximum correlation coefficient value between tf and document length is 0.2123 for Wt2g collection that means the relationship between two parameters in the collections is weak (see Table 4).

[1] Commonwealth Scientific & Industrial Research Organization

Table 4. Maximum correlation coefficient values between tf and document length

Wt2g	*Wt10g*
0.2123	0.1747

A simple analysis of this result shows that a high value of term frequency in a document cannot show much information about the length of the related document. Distinctively, the high frequency of a term in a document is not a good indicator to estimate that the term is located in a long document. Indeed, using document length parameter for term frequency normalization reduces the terms weights of a long document in according with its length. But, the retrieval chances of small documents containing the low frequency terms will be increased.

Afterward, the correlations between term frequency and document frequency of terms in 50 documents are calculated randomly whose maximum values are shown in Table 5. The maximum correlation coefficient values (0.2425 for Wt10g) display a low correlation between two parameters.

Table 5. Maximum correlation coefficient values between tf and df

Wt2g	*Wt10g*
0.1878	0.2425

Therefore, a high value of term frequency in documents cannot indicate its frequency in the whole collection. This means that high frequency of a term is not a good indicator for estimating that the term is situated in a large number of documents.

In other word, all of the terms within a long document are not good indicators for document content and it is appropriate to consider more significant terms for document discrimination. Hence, it is important to determine significant document terms and they should be relied much more than the other terms in term-weighing schemes. In the next section, the analyses are continued with declaration a novel factor for document length normalization.

5. TERM IMPORTANCE DEGREE FACTOR

The first concept which can be driven from previous results is that high frequency of a term in a document does not necessarily specify the scale of document length and document frequency. We considered the term importance degree within document (tid) which is determined by ranking the terms based on term frequency within each document as a significant factor to apply in term-weighting schemes. This factor assesses the importance of document terms not only by frequency, but also by frequency rank. The notations which are used are as follow:

d: a document
t: a term
q: a *query*
tf: frequency of term t in document d
df: number of documents containing term t
dl: length of document d

It must be regarded, the fewer value of tid, the higher importance degree of term t in document d. A term t, which repeated 5 times in a document containing 50 terms, and 5 times in another document with the same length, may not have the same importance in these documents. Thus, the frequency of term in a document should be compared with the frequencies of other terms in the same document. In addition, an important term, but not frequent in a long document must not be penalized for the reason that it is in a long document with numerous terms.

To evaluate the impact of tid factor on discrimination the terms, its correlation coefficient with term-weighting parameters have been calculated. It consists of calculation the couple correlation (tid, tf), (tid, idf) and (tid, dl). Next, two new parameters called "Balanced term importance degree" and "Average term importance degree" based on tid factor have been introduced. These parameters are proposed some modifications of Okapi term-weighting function.

5.1 Correlation Between tid and Term-weighting Parameters

The objective of this section is to measure the impact of tid on term discrimination. The terms and documents which have been used for calculating the correlations are the same as the terms that used in previous experiment. Table 6 shows the obtained maximum correlation coefficient values between tid and tf.

Table 6. Maximum correlation coefficient values between tid and tf

Wt2g	*Wt10g*
-0.3306	-0.1261

The obtained values don’t signify a high dependency between these parameters. It means that the rank of terms is not comparable with each other in different documents. Hence, a term with high value of term frequency in a document may have low importance degree. In other word, high frequency of a term is not a good indicator for importance degree factor. Next, the correlations between tid and document frequency parameters were calculated whose results are presented in Table 7.

Table 7. Maximum correlation coefficient values between tid and df

Wt2g	*Wt10g*
-0.1478	-0.2425

The maximum correlation values of table 7 indicate a low and inverse dependency between two parameters. It means that high df value of a term in a collection does not present a high importance degree in the documents. In combination with precedent section, it is realized that tf and df values are not significant indicators for term importance degree factor.

Finally, the correlation coefficients between tid and document length parameters were calculated. The maximum values of this correlation in two test collections are presented in Table 8.

Table 8. Maximum correlation coefficient values between tid and document length

Wt2g	*Wt10g*
0.8940	0.9101

These values show that the correlation between document length and tid factor is higher than the correlation between tid and the other parameters. Therefore, the influence of document length on tid can be verified for terms discrimination in weighting schemes. It is assumed that, while the length of a document increases, the importance degrees of the terms within document decrease. In other word, all of the terms within a long document are not good indicators for document content and significant terms must be more considered for retrieval. This can be done in the indexing phase that makes to reduce the space of indexing terms. Also, it can be considered in retrieval evaluation phase that can ameliorate the efficiency of research. In the next sections, tid factor is used in term-weighting scheme by two new proposed parameters.

6. BALANCED TERMS IMPORTANCE DEGREE

Based on the previous analysis, there is a significant correlation between document length and term importance degree factor. First question was that, how the term importance degree factor can be applied in term-weighting schemes and how change this quantitative factor to a qualitative factor. We have introduced the B_Rank parameter refers to "Balanced term importance degree" which indicates percentage of top terms of documents ranked by their frequencies. The value of B_Rank parameter revises the interaction between frequency and term importance degrees in the documents. This means that a document is more appropriate for a query term, while the query term is one of the major terms in document.

Therefore, top terms value is boosted by introducing the β factor in weighing scheme is as follow:

$$W = \frac{tf \times \beta \times log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right)}{2\times\left(0.25 + 0.75\,\frac{dl_j}{Avgdl}\right) + tf_{ij}} \qquad (5)$$

Where:
β: if $tid_{ij} \leq$ #B_Rank(d_j) then β = α else β = 1
tid_{ij}: importance degree of term *i* within document *j*
B_Rank(d_j): percentage of ranked terms based on term frequency in document *j*
α: a constant which is determined empirically

This means that certain terms (identified by B_Rank) are weighted more than the other terms. Following, the impact of B_Rank parameter on the weighting scheme performance is verified.

6.1 Evaluation of B_Rank Impact

To evaluate the validity of proposed weighting model (formula 5), it is conducted several runs on the test collections. Some details and the percentage of average precision improvement are presented in Tables 9 and 10. Several values of α (2, 3, 4 and 5) and B_Rank parameter are considered to verify the effectiveness of tid factor. Baseline values show the results of MAP with formula 1.

Table 9. MAP in Wt2g Collection (Baseline and different settings)

B_Rank = 5%				
Baseline	α = 2	α = 3	α = 4	α = 5
0.2635	0.2872	0.2900	0.2915	0.2915
B_Rank = 10%				
0.2635	0.2919	0.3010	0.2989	0.305
B_Rank = 20%				
0.2635	0.2920	0.2925	0.2912	0.2912

Table 10. MAP in Wt10g Collection (Baseline and different settings)

B_Rank = 5%				
Baseline	α = 2	α = 3	α = 4	α = 5
0.1884	0.1955	0.2012	0.2015	0.1965
B_Rank = 10%				
0.1884	0.1982	0.2213	0.200	0.200
B_Rank = 20%				
0.1884	0.2103	0.2174	0.2190	0.2070

The results show the positive impact of B_Rank parameter on retrieval precisions. Also, the best results especially on average precision were obtained with B_Rank = 5% to 10% and α = 3. These values mean that only 5% to 10% of important terms in a document are principal factor to show its relevance for a query.

7. AVERAGE TERMS IMPORTANCE DEGREE

It is observed that document length parameter plays an important role in term frequency normalization. Beside, distribution of documents according to their lengths in larger collections is more widespread. Our second proposed parameter called "Avg _Rank", which uses the tid factor for normalization term frequency. This parameter represents the average of importance degree of document terms and is defined as follows:

$$Avg\_Rank_j = \frac{\sum_{i=1}^{|d_j|} tid_{ij}}{|d_j|} \qquad (6)$$

Where:
Avg_Rank$_j$: average of terms importance degree of document *j*
$|d_j|$: # unique terms in document *j*

The study of Avg _Rank parameter impact on BM25 formula is as follow:

$$W = \frac{tf \times log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right)}{\left(c_1 + c_2 \times Avg\ \_Rank_j\right) + tf_{ij}} \qquad (7)$$

Where:
$c_1 = 0.1$, $c_2 = 0.25$

The main advantage of Avg_Rank contrary to document length parameter is that, it is unique in each document and independent from other documents. The intuition behind this setting was that, using term importance degree has a positive impact in term-weighting scheme and it will be useful to establish a specific term frequency normalization parameter for each document. The small value of Avg_Rank in a document means that the majority of document terms are important and this parameter must increase the weight of these terms. Contrarily, if the value of this parameter is high, it means that the document has a lot of duplicated terms and therefore document' score will be reduced. The average precisions obtained with formula 7 on the test collections are presented in Fig. 1.

As it is seen in Fig. 1, the average precisions have been increased as high as +11.7 % and +17.5% for Wt2g and Wt10g collections respectively. It can be concluded that the proposed schemes based on term importance degree factor produce better discrimination between sizes of documents in relation to the document length parameter which is used in many weighting schemes.

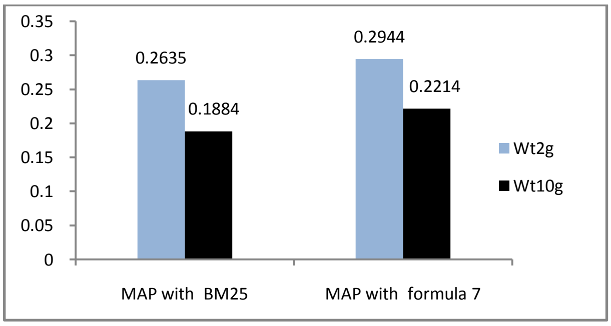

Fig.1. MAP using BM25 and formula7 in Wt2g and Wt10g

8. CONCLUSION

The proposal of this work was a factor named "Term importance degree" related the rank of terms in the documents. The aim was to provide an auxiliary factor for normalization frequency of term and document length parameters, which plays the principal roles in retrieval process especially in large and heterogeneous collections. This factor is proposed to improve the impact of both parameters. Based on this factor, two new parameters were defined: B_Rank and Avg_Rank. The preparation of these parameters was supported by an original approach based on statistical study of correlations and performance of various parameters within presented weighting scheme.

The intuition behind the presented approach was that the terms contained in a document can be rearranged in order to increase their relevance for retrieval. The experiments on TREC collections have revealed that the proposed schemes have effective performance and assess the variety of term frequencies and document sizes more effectively.

9. REFERENCES

[1] Amati, G. & van Rijsbergen, C. J., Probabilistic models of information retrieval based on measuring the divergence from randomness. In ACM Transactions on Information Systems (TOIS), volume 20(4), pages 357 - 389, 2002.

[2] Anh, V. & Moffat, A., Simplified similarity scoring using term ranks, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 2005, Salvador, Brazil.

[3] Baeza-Yates, R., & Ribeiro-Neto, B., Modern information retrieval. Harlow, England: Addison - Wesley Longman Ltd, 1999.

[4] Buckley, C., Singhal, A., Mitra M. & Salton, G. (1996). New retrieval approaches using SMART. In Proceedings of TREC-4, (pp. 25-48), Gaithersburg, MD: NIST Publication #500-236.

[5] Buckley C. & Voorhees E.M., Evaluating evaluation measure stability. In Proceedings of the 23rd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, pages 33–40, ACM Press, 2000.

[6] Craswell, N. & Hawking D., Overview of the trec-2002 web track. In The 11th Text Retrieval Conference, TREC'2002, pages Gaithersburg, Maryland, USA, NIST Special Publication SP 500-251, 2002.

[7] Cummins, R. & O'Riordan, C., An evaluation of evolved term-weighting schemes in information retrieval. In CIKM'05: Proceedings of the 14th ACM international conference on Information and knowledge management, pages 305-306, New York, NY, USA, 2005, ACM Press.

[8] Fang, H., Tao, T. & Zhai, C., A formal study of information retrieval heuristics. SIGIR 2004: 49-56.

[9] Luhn, H. P., The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2 (2), p. 159-165 and 317, April 1958.

[10] Maron, M., Automatic indexing: an experimental enquiry. Journal of the ACM, 24(8): 404-417, 1961.

[11] Robertson, S. E & Sparck Jones, K., Relevance weighting of search terms. Journal of the American Society for Information Science, 27: 129 - 146, 1976.

[12] Robertson, S., Walker, S., M. M. Beaulieu, Gatford, M. & A. Payne, Okapi at trec-4. In NIST Special Publication 500-236: The Fourth Text Retrieval Conference (TREC-4), pages 73 - 96, 1995.

[13] Robertson, S. E. & Walker, S., Okapi/Keenbow at TREC-8. In E M Voorhees and D K Harman, editors, The Eighth Text Retrieval Conference (TREC-8), pages 151- 162. Gaithersburg, MD: NIST, 2000, NIST Special Publication 500-246.

[14] Salton, G. & McGill, M.J., Introduction to Modern Information Retrieval. McGraw-Hill, New York 1983.

[15] Salton, G., Syntactic approaches to automatic book indexing. In Proc of the annual meeting on Association for Computational Linguistics (ACL) (1988), pages 204-210, Department of Computer Science, Cornell University, Ithaca, New York, 1988.

[16] Salton, G. & Buckley, C., Term-Weighting Approaches in Automatic Text Retrieval, Information Processing & Management, 24(5), pp. 513-523, 1988.

[17] Singhal, A., Buckley, C. & Mitra, M., Pivoted document length normalization. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21–29, 1996.

[18] Singhal, A., Salton, G., Mitra M. & Buckley C. (1996), 'Document length normalization'. Information Processing & Management 32, 619–633.

[19] Singhal, A., Choi, J., Hindle, D., Lewis, D.D. & Pereira, F., (1999). AT&T at TREC-7, In Proceedings of TREC-7, (pp. 239-251), Gaithersburg, MD: NIST Publication #500-242.