TÜBİTAK

Turk J Elec Eng & Comp Sci, Vol.19, No.6, 2011, © TÜBİTAK
doi:10.3906/elk-1003-448

Investigation of Luhn's claim on information retrieval

İlker KOCABAŞ[1,*], Bekir Taner DİNÇER[2], Bahar KARAOĞLAN[1]
[1] *International Computer Institute, Ege University, 35100 İzmir-TURKEY*
e-mails: {ilker.kocabas,bahar.karaoglan}@ege.edu.tr
[2] *Department of Statistics, Muğla University, 48100 Muğla-TURKEY*
e-mail: dtaner@mu.edutr

Received: 03.05.2010

Abstract

In this study, we show how Luhn's claim about the degree of importance of a word in a document can be related to information retrieval. His basic idea is transformed into z-scores as the weights of terms for the purpose of modeling term frequency (tf) within documents. The Luhn-based models represented in this paper are considered as the TF component of proposed TF × IDF weighing schemes. Moreover, the final term weighting functions appropriate for the TF × IDF weighting scheme are applied to TREC-6, -7, and -8 databases. The experimental results show relevance to Luhn's claim by having high mean average precision (MAP) for the terms with frequencies around the mean frequency of terms within a document. On the other hand, the weighting, which significantly discriminates the importance between low/high frequencies and medium frequencies, degrades the retrieval performance. Therefore, any weighting scheme (TF) that is directly proportional to tf has a probability of high retrieval performance, if this can optimally indicate the difference of the importance regarding tf values and also optimally eliminate the terms that have high frequencies.

Key Words: *Luhn, information retrieval, term weighting, indexing*

1. Introduction

Basically, the context of a text is formed by the total semantics gathered from all terms within the text; a term may be a letter, a number, a character, or any combination of these that has a meaning. Different terms constituting a text do not contribute the same amount to the semantic information transferred via the text. That is to say, the importance of each term may vary due to its contribution. If this variation in representing semantic information can be reflected by means of weights given to index terms, the information content of a text can be characterized more precisely [1].

An approach to term weighting was originally conceived by Luhn [2]. He proposed that each word can be weighted by its relative frequency with respect to all the words of a given text. Although this was the first time "term frequency" was used for weighting index terms, Luhn described the relation between the frequency of a

*Corresponding author: International Computer Institute, Ege University, 35100 İzmir-TURKEY

term (tf) within a text or document and the informative content or significance of a term within this document in his later work [3]. His claim is represented graphically in Figure 1 as the plot of the term frequencies with respect to their level of importance. This relation can further be explained as follows:

a) The terms with medium frequencies are more important than the terms that have low or high frequencies. Rare words with low frequencies that are below the lower cut-off C_L and the common words with frequencies exceeding the upper cut-off C_u do not contribute significantly to the content of the text.

b) The resolving power, or the degree of strength in discriminating the content, of significant words in a text reaches its peak at a point within the medium frequencies range (between the 2 cut-offs) and falls in both directions, becoming almost negligible at the cut-off points.

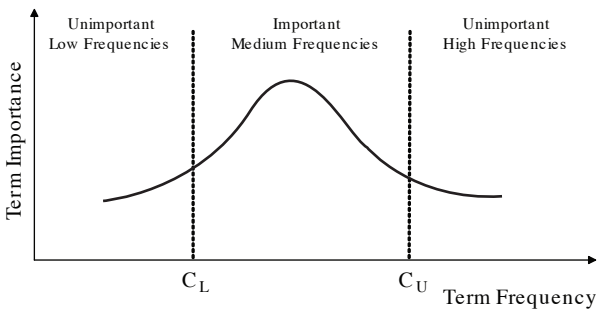

Figure 1. Relation between term importance and term frequency [3, adaptive].

Even though Luhn put forth his claim just for the text summarization problem by sentence selection, by looking at the promising results we think that the most exciting point is that it can be transferred completely to the field of information retrieval (IR) for constructing indexing models.

The modeling of *tf*, which is the way of expressing the degree of the importance or contribution of a term to the document context, is regarded as the TF component of a basic TF $\times$ IDF weighting scheme. Salton [4] and Minker et al. [5] showed experimentally that using such a TF component in weighting index terms resulted in superior retrieval performance over unweighted terms. Moreover, Robertson et al. proposed alternative TF schemes [6,7]. Although these studies were inspired from the idea of taking *tf* into account for weighting, which was mentioned before by Luhn, they are differentiated from Luhn's claim by the assumption that the weighting of a term (TF) is directly proportional to the frequency of a term (tf). Briefly, this assumption is the general approach under the weighting models, henceforth defining explicitly the contribution of a term to the content of a document. Using a stop-word list is one of the means of eliminating some unimportant terms that are of high frequency. However, the gap between Luhn's description of degree of significance, as interpreted in point (b) above, and these kinds of assumptions still exists.

One of the different weighting approaches, based on combining the interdocument and intradocument term frequencies, is the term discrimination value (TDV) approach [8]. Salton et al. [9], examining the behaviors of TDVs, noted the impact of the frequency distribution of each term within the collection, in addition to Luhn's viewpoint. The given conclusions were basically an expansion of Luhn's claim collection-wide. On the other hand, TDV, which is primarily intended to discriminate the vocabulary terms of a collection, performs the same task intended by interdocument frequency [10] (IDF). Hence, it is impossible to reach the judgment that TDV has the same extensions as the term significance interpretation of Luhn within the document boundary.

There are alternative approaches to automatic indexing based on a probabilistic viewpoint. There are many models for this approach, including the works of Harter [11,12], Robertson and Sparck Jones[13], Cooper and Maron [14], Croft and Harper [15], Robertson et al. [16], Fuhr [17], Turtle and Croft [18], Wong and Yao [19], Ponte and Croft [20], Hiemstra and de Vries [21], and Amati and van Rijsbergen [22]. These approaches are primarily concerned with the estimation of some model parameter(s) and the constructing of a weighting formula toward integration of these parameters. Such weighting formulas are complex and it is very hard to conclude that they take into account Luhn's viewpoint on the importance of words. On the other hand, we may say that there is no explicit evidence of Luhn's claim considered as a whole in these approaches.

To summarize all of the discussions above, we may say that the available approaches that explain the contribution of a term in a document's context with the notion of the term frequency deviate slightly from Luhn's viewpoint. Luhn assessed the frequencies of terms relative to the average frequency of all terms occurring in that document instead of taking into account the pure frequencies of terms. To the best of our knowledge, we can say that the validity of Luhn's claim from the perspective of IR has not been investigated completely yet, in spite of the fact that it is the pioneering theoretical foundation that refers to the relationship between terms/words and semantic information via the notion of term frequency.

In this study, in order to express Luhn's viewpoint completely as a weighting parameter, we constructed 2 different TF functions combined with an IDF component, representing Luhn-based TF $\times$ IDF schemes. We then tested their effectiveness in IR with TREC databases. In the following sections, our term weighting models/functions are explained precisely and experimental results are given.

2. Notation

Any set of documents may be represented as a term $\times$ document matrix X, shown in Figure 2. Rows and columns of matrix X represent t_i $(i = 1 \ldots r)$ terms and d_j $(j = 1 \ldots c)$ documents, respectively. Each cell of X, x_{ij}, indicates the number of occurrences of term t_i in document d_j.

$$X = \begin{array}{c|c||cccccccc} & & \multicolumn{8}{c}{Documents} \\ & & d_1 & d_2 & d_3 & \square & d_j & \square & d_c \\ \hline\hline Terms & t_1 & x_{11} & x_{12} & x_{13} & \square & x_{1j} & \square & x_{1c} \\ & t_2 & x_{21} & x_{22} & x_{23} & \square & x_{2j} & \square & x_{2c} \\ & t_3 & x_{31} & x_{32} & x_{33} & \square & x_{3j} & \square & x_{3c} \\ & \square & \square & \square & \square & \square & \square & \square & \square & \square \\ & t_i & x_{i1} & x_{i2} & x_{i3} & \square & x_{ij} & \square & x_{ic} \\ & \square & \square & \square & \square & \square & \square & \square & \square & \square \\ & t_r & x_{r1} & x_{r2} & x_{r3} & \square & x_{rj} & \square & x_{rc} \end{array}$$

Figure 2. Term $\times$ document matrix.

3. Luhn-based TF $\times$ IDF models

Luhn's qualitative viewpoint, explained in section 1, can be expressed quantitatively by means of z-scores, as follows:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{s_j^2}}. \tag{1}$$

Here, $\overline{x}_j$ indicates the mean of the term frequencies in document d_j. For any term t_i in document d_j, the z-score gives the difference of term frequency, which is represented by x_{ij}, from the mean frequency value of the terms in d_j, which is $\overline{x}_j$, with respect to intradocument standard deviation s_j. This is to standardize the deviation of frequencies of terms from the mean frequency with respect to standard deviation: standardized representation for occurrences of term t_i in document d_j. At this point, standardization refers to being in the same scale. Consequently, the z-score of each term can be compared among all of the documents in which it occurred, independently of the document length. In fact, this process is also referred to as normalization, which is one of the basic efforts performed in the available indexing methods.

It is possible to restate the relation between term importance and frequency given in Figure 1 as in Figure 3 with respect to the z-transform of term frequencies. According to this transformation, the term that has a value of $z = \alpha$ is assumed to be the most important term related to the semantic information.

That is to say, the z-score of the frequency value that semantically indicates the most important term for a document in Figure 3 is equal to α.

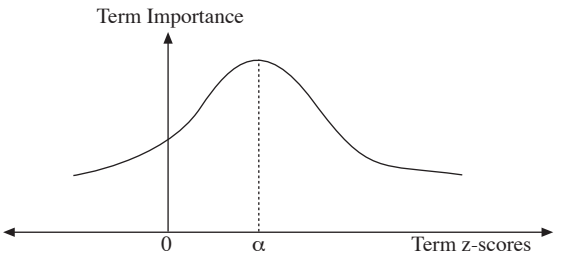

Figure 3. Term importance vs. term z-scores.

For any term t_i in document d_j, z_{ij} (the $z-$score of t_i) and the term importance (the weighting TF_{ij} of t_i) are inversely related to each other. This inverse relation may be expressed quantitatively as 2 different functions, which are shown in Eqs. (2a) and (2b).

$$TF1 = \frac{1}{|Z_\alpha| + 1},\ Z_\alpha = (\alpha - z_{ij}) \tag{2a}$$

$$TF2 = \frac{1}{Z_\alpha^2 + 1},\ Z_\alpha = (\alpha - z_{ij}) \tag{2b}$$

The weighting functions given in Eqs. (2a) and (2b) are named as *TF1* and *TF2*, respectively. Both functions generate values in the interval of [0,1], where the maximum 1 is at $Z_\alpha = 0$; that is, equal to $z_{ij} = \alpha$. By equation (2b), the importance of a term that has a $z-$score in the interval $[\alpha - 1, \alpha + 1]$ is more valuable, and otherwise it is less than the calculation of equation (2a). Assume 2 terms, t_1 and t_2, with Z_α-scores equal to z_1 and z_2, respectively, where z_1 is in the interval [-1,1] and z_2 is out of the interval [-1,1], which is also out of the mid-frequencies region. The importance of t_1, which is calculated by *TF2*, is higher than the value calculated by *TF1*, whereas t_2 has a lesser importance in the event of using *TF2*. In other words, *TF2* is more condensed for determining the mid-frequency or mid-frequencies region, which Luhn claimed as the important

term(s). The actual importance values that can be calculated from functions *TF1* and *TF2* are given in Figure 4 with respect to the Z_α-scores.

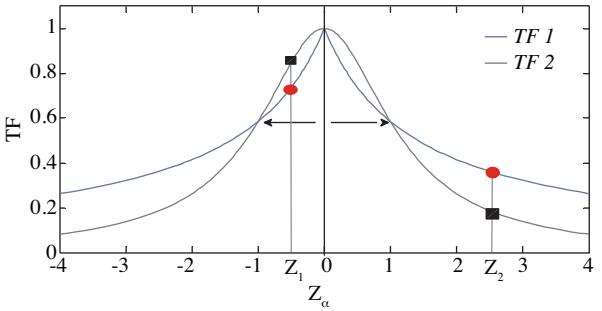

Figure 4. TF vs. Z_α-scores.

Under the assumption of mutually exclusive behavior among terms, the ranking function, which measures the degree of relevancy of a document with respect to a query, equals the multiplication of the weights of the terms included. It is also possible to define the ranking function in the additive form of weights by using a logarithm transformation.[1]

$$TF - score(d_j) = \sum^{\text{for } \forall t_i \text{ inquery}} \log_2 (TF_{ij} + 1) \tag{3}$$

After adding the IDF component, the ranking function will be expressed as:

$$score(d_j) = \sum^{\text{for } \forall t_i \text{ in query}} \log_2 (TF_{ij} + 1) \times IDF. \tag{4}$$

Sparck Jones' *idf* [10] was used for the IDF component of the TF × IDF schemes: $idf = \log_2(N/n_o + 1)$, where N is the total number of documents in the collection and n_o is the number of documents in which a term (t_i) is observed. Consequently, the final term weighting functions are given in Eq. 5; they are named as *WTF1* and *WTF2*, in accordance to using *TF1* and *TF2* as the TF components.

$$WTF1 = \log_2 \left(\frac{1}{|Z_\alpha| + 1} + 1 \right) \cdot idf, Z_\alpha = (\alpha - z_{ij}) \tag{5a}$$

$$WTF2 = \log_2 \left(\frac{1}{Z_\alpha^2 + 1} + 1 \right) \cdot idf, Z_\alpha = (\alpha - z_{ij}) \tag{5b}$$

4. Empirical study

The main objectives of the experimental studies can be summarized as:

- Validating the relation that is modeled based on Luhn's claim.
  - Investigating the IR performance of the represented weighting schemes.

[1] $\log(\prod a_i) = \sum \log(a_i)$

In order to test whether Luhn's quantitive description of term importance meets the purpose of information retrieval applications, we tried to find an answer to the following questions:

i. Is there any mid-frequency or range of mid-frequencies that indicate the high importance of term(s), as Luhn claimed for intradocument characteristics?

ii. If yes, is it possible to perform effective retrieval (using a Luhn-based TF in a TF $\times$ IDF scheme that includes interdocument statistics) by assigning weights to terms based on Luhn's interpretation?

iii. For the purpose of good retrieval performance of the TF $\times$ IDF weighting scheme, what are the characteristics that would be needed for an optimal model of *tf* in the document boundary?

The following subsections investigate the term importance problem in light of these aspects.

4.1. Experimental setup

We carried out all of our experiments on TERRIER[2] (TExt RetRIevER) platform. A single-pass indexer was used for indexing. The built-in matching model was changed with the proposed weighting functions.

We used the test collection of TREC (Text Retrieval Conference) from disks 4 and 5. For this collection, we performed tests on each of the 50 topics in TREC-6, TREC-7, and TREC-8. The TREC-6 test collection consists of about 2.1 GB of data with about 556,000 documents, from the Congressional Record (CR), Financial Register (FR), Foreign Broadcast Information Service (FBIS), and Los Angeles Times (LA) collections. In TREC-7 and TREC-8, the CR collection, which includes large-size documents, was removed from the indexing. After that, the average length of the documents decreased from 557 tokens to 512 tokens.

Each of the topics has the same structure, consisting of 3 fields. These are the title, which includes the most related words (1-3 words); the description, which gives a wider explanation about the query (1 or 2 sentences); and a narrative, which contains specific conditions on accepting or rejecting documents (a paragraph). In our experiments, we used only the title field. In the indexing phase, Porter's stemming algorithm [23] was used, but we did not use a stop list for the purpose of protecting (or not damaging) the natural statistics of the documents. On the contrary, in the retrieving phase, we used a stop list of 733 words in the queries.

4.2. Estimation of α based on TF component

The weighting schemes for the TF component presented in this paper are only dependant on document context. That is to say, there is no component regarding the potential information of a word, or information carried by a word to be free from context, like the IDF component of TF $\times$ IDF schemes. Therefore, for these models that approach the terms as equally likely, it is the best to use only the title of the topics in order to determine the most appropriate α value or to find out if a relation, which is claimed by Luhn, exists between the query and the document.

In these experiments, the retrieval performances of models *TF1* and *TF2* were measured for 16 different values of α from 0 to 3 with a step of 0.2 ($\alpha = 0, 0.2, 0.4 \ldots 2.8, 3.0$). The $z-$value reaches its peak at the mean frequency of the document when α is zero. Thus, "frequency at peak" $<$ "mean frequency" for $\alpha < 0$. These are below the peak frequency value, such that we found poor performances for $\alpha < 0$. In other words, we only give the performances for $\alpha \geq 0$ because of poor results obtained at $\alpha < 0$. Moreover, mean average precision

[2]TERRIER home page, http://ir.dcs.gla.ac.uk/terrier/

(MAP) was used for the retrieval performance measure. The tests results for TREC-6 through TREC-8 are shown in Figures 5, 6, and 7, respectively.

For the TREC-6 ad hoc retrieval task (given in Figure 5), both models had the same mean average precision (MAP = 0.09) at $\alpha = 0$. The performance of model *TF2* increased until α reached 1.0 (max MAP $\cong$ 0.14). The same increase was observed with *TF1* with a different peak value of α ($\alpha = 1.4$) and a higher performance (max MAP $\cong$ 0.16). After these values, the performances of both models started to decline. The performance decline was steeper for *TF2*.

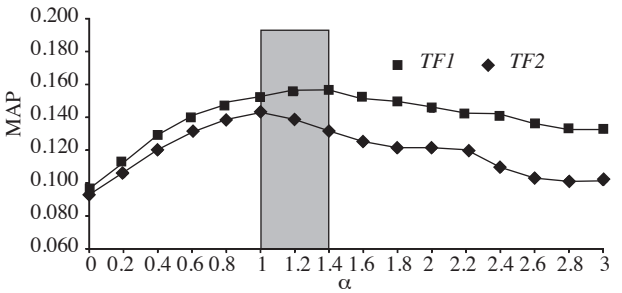

Figure 5. α vs. MAP for disks 4 and 5 of TREC-6, topics 301-350, title only.

For the TREC-7 ad hoc retrieval task (given in Figure 6), both models had nearly the same MAP value ($\cong$0.06) at $\alpha = 0$. However, the performance of *TF2* increased until α reached 0.8 and declined beyond this value; *TF1* retrieval performance increased until α reached 1.0 and remained almost constant afterwards.

The TREC-8 test results (given in Figure 7) look graphically like the TREC-6 test results. That is to say, the observation of performance increase and decrease with respect to the value of α was significant. The peak MAP values obtained for *TF1* and *TF2* were at $\alpha = 0.8$ and at $\alpha = 0.8$, respectively.

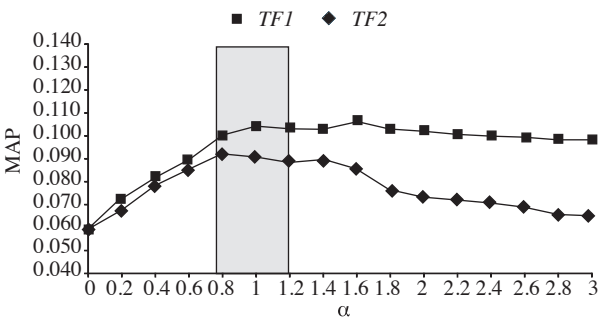

Figure 6. α vs. MAP for disks 4 and 5 without CR of TREC-7, topics 351-400, title only.

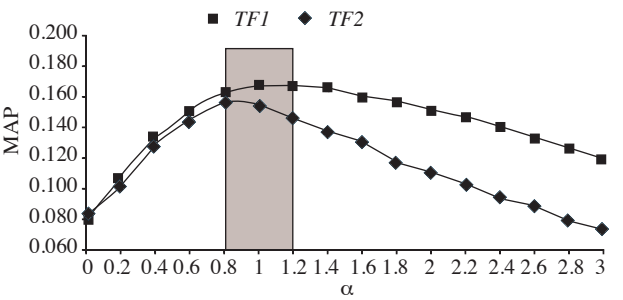

Figure 7. α vs. MAP for disks 4 and 5 without CR of TREC-8, topics 301-350, title only.

As we took each peak value, or highest performance test result, of both models for TREC-6, it was suitable to estimate the optimal value of α in the interval of [1.0,1.4]. By the same approach, the intervals of TREC-7 and TREC-8 were nearly same, approximately equal to [0.8,1.2]. The difference of the intervals observed from the TREC-6 tests and the TREC-7 and TREC-8 tests may be based on the insufficiency of standardization of z-scores due to the removing of longer CR documents. Despite the deficiencies of this standardization, we may choose $\alpha = 1$ as the optimal value.

The same situation described above for TREC-6 remained true for TREC-7 and TREC-8. In other words, increasing or decreasing the value of α from the optimal value (the interval of the existing maximum value)

led to performance degradation for TREC-7 and TREC-8. This finding was more obvious for TREC-8 and less obvious for TREC-7.

4.3. Experiments on Luhn-based TF-IDF schemes

Models constructed on Luhn-based TF-IDF schemes were explained in the previous section. The weighting functions of such schemes were named as *WTF1* and *WTF2* with respect to the TF components used. *WTF1(*α*)* and *WTF2(*α*)* represent those functions varying with respect to parameter alpha (α). The developed TF-IDF models were run on TREC-6, TREC-7, and TREC-8 datasets for values of $\alpha = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$.

Mean average precision (MAP) and R-precision were used as retrieval performance measures. The representation of performance measures was made by plotting MAP and R-precision values of the models on each dataset. We compared our models' performance results to those of Okapi BM25 [24], a broadly used weighting function.

The test results of the TREC-6 ad hoc retrieval task dataset are given in Figure 8. The performance of model *WTF2* shows the same *TF2* characteristics (subsection 4.2) such that its performance increased until α reached 1.0, and decreased after that point. The highest performance measures obtained by *WTF2* were 0.2094 and 0.1709 for R-precision and MAP, respectively, at the value of $\alpha = 1.0$, and its MAP was approximately 20% higher than the highest MAP value of *TF2* (1.0), which was measured at approximately 0.14. On the other hand, the *WTF1* performance exhibited some characteristics different from the *TF1* performance, such that *WTF1* reached its peak MAP value ($\cong$0.185) at $\alpha = 2.0$ and peak R-precision value ($\cong$0.229) at $\alpha = 1.5$. Even if the retrieval performances of *TF1* decreased while the value of alpha exceeded these peak points, the MAP value in particular became nearly stabilized at about MAP $\cong$ 0.18. The highest MAP obtained from *WTF1* was also approximately 15% higher than the highest MAP value of *TF1* (1.4), which was measured at approximately 0.16. Finally, *WTF1* and *WTF2* had lower R-precision and MAP values than the respective values of BM25 (0.255, 0.206); approximately, both measures were 11% lower for *WTF1* and 20% lower for *WTF2*.

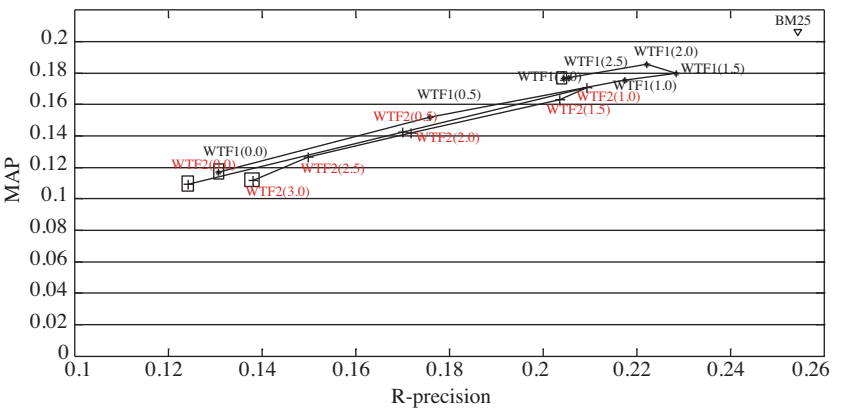

Figure 8. MAP vs. R-precision for disks 4 and 5 of TREC-6, topics 301-350, title only.

The other test results obtained from the TREC-7 ad hoc retrieval task dataset are given in Figure 9. Parallel to the TREC-6 results, the performance of model *WTF2* also showed the same *TF2* characteristics (subsection 4.2) over the TREC-7 dataset, such that its performance increased until α reached 1.0, and decreased after that point. The highest performance measures obtained by *WTF2* were 0.1779 and 0.1309 for R-precision

and MAP, respectively, at the value of $\alpha = 1.0$; its MAP was approximately 45% higher than the highest MAP value of *TF2* (1.0), which was measured at approximately 0.09. On the other hand, the *WTF1* performance exhibited some characteristics different from the *TF1* performance, such that *WTF1* reached its peak MAP value ($\cong$0.153) and peak R-precision value ($\cong$0.204) at $\alpha = 2.5$. Even if the retrieval performance of *WTF1* decreased while the value of alpha exceeded this peak point, the performance measures became nearly stabilized around MAP $\cong$ 0.15 for values of α greater than 1.5. The highest MAP obtained from *WTF1* was also approximately 50% higher than the highest MAP value of *TF1* (1.0), which was measured at approximately 0.1. Finally, *WTF1* and *WTF2* had lower R-precision and MAP values than the respective values (0.204, 0.153) of BM25; approximately, both measures were 5%-7% lower for *WTF1* and 20%-25% lower for *WTF2*.

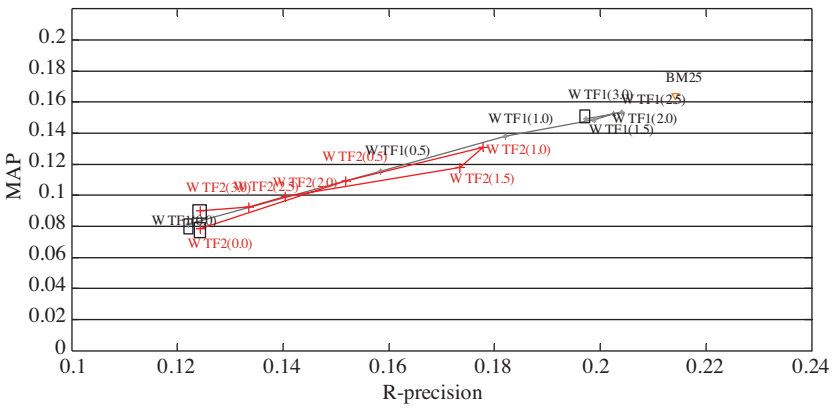

Figure 9. MAP vs. R-precision for disks 4 and 5 of TREC-7, topics 301-350, title only.

The final test results obtained from the TREC-8 ad hoc retrieval task dataset are given in Figure 10. The performance of model *WTF2* showed the same *TF2* characteristics (subsection 4.2), such that its performance increased until α reached 1.0, and decreased after that point. The highest performance measures obtained by *WTF2* were 0.2401 and 0.1902 for R-precision and MAP, respectively, at the value of $\alpha = 1.0$; its MAP was approximately 22% higher than the highest MAP value of *TF2* (1.0), which was measured at approximately 0.155. On the other hand, the *WTF1* performance exhibited some characteristics different from the *TF1* performance, such that *WTF1* reached its peak MAP value ($\cong$0.202) and peak R-precision value ($\cong$0.258) at $\alpha = 2.0$. Even if the retrieval performance of *WTF1* decreased while the value of alpha exceeded this peak point, the MAP value became nearly stabilized at MAP $\cong$ 0.2 for the α values in the interval of [1.0,2.5]. Contrary to other tests, a significant performance fall was observed at $\alpha = 3.0$. The highest MAP obtained from *WTF1* was also approximately 20% higher than the highest MAP value of *TF1* (1.4), which was measured at approximately 0.167. Finally, *WTF1* and *WTF2* had lower R-precision and MAP values than the values of BM25 (0.278, 0.219); approximately, both measures were 8%-9% lower for *WTF1* and 15% lower for *WTF2*.

Generally speaking, the optimal mid-frequency point found at $\alpha = 1.0$ from the previous experiments remained the same for the intended mid-frequencies version of *WTF2*; however, for *WTF1*, the optimal alpha, which indicates the most important term frequency, was observed to be higher than 1.0, varying between 1.5 and 2.5 according to test datasets. One of the other findings was that the retrieval performance increased in the event of including IDF, a measure based on interdocument statistics, in the proposed TF components, measurements based on intradocument statistics. Performance advances for *WTF1* and *WTF2* were approximately 20% for the TREC-6 and TREC-7 datasets; meanwhile, they were 40%-50% for TREC-7. On the other hand, the overall performances of *WTF1* and *WTF2* were approximately 10% and 20% lower than BM25.

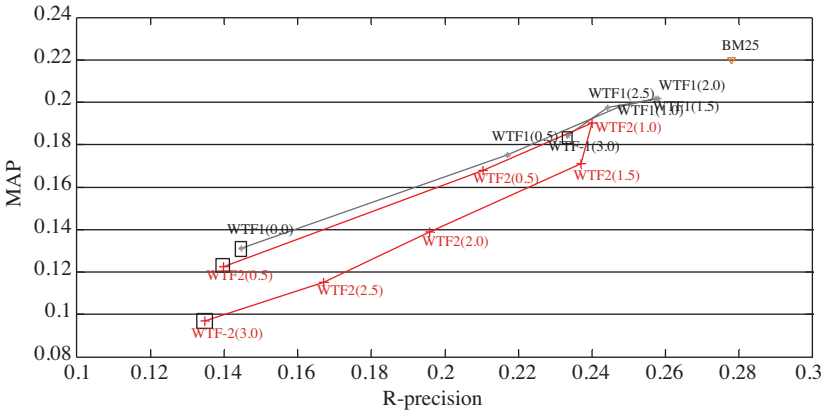

Figure 10. MAP vs. R-precision for disks 4 and 5 of TREC-8, topics 301-350, title only.

5. Conclusions

In this study, we formulated 2 different TF weighting functions based on Luhn's claim about the importance of words within a document. These TF functions were located in 2 different document ranking functions applicable to a basic TF $\times$ IDF scheme. Our primary goal was to investigate the validity of Luhn's viewpoint for IR.

All of the experiments were carried out on TREC-6 through TREC-8 ad hoc tracks. Experimental results obtained from using the TF-only weighting functions showed that there was an optimal value for α approximately equal to 1.0. Moreover, the decline of performance in both directions from this optimal value also strengthened the validity of Luhn's claim for IR. Consequently, these findings are indicators of the existence of a region that includes important terms, or mid-frequencies, in Luhn's claim.

In addition to the represented Luhn-based TF components, use of IDF components advances the retrieval performance significantly; these weighting functions in the TF $\times$ IDF scheme are called final weightings. On the other hand, the experimental results of such final weightings showed some characteristics different than those of the TF-only cases. Although the optimal value of alpha remained unchanged for the *WTF2* function that consisted of *TF2* (condensed to indicate the region of higher importance terms), the location of the most important term for *WTF1* increased (higher value of alpha) and the retrieval performance remained nearly stable for higher alpha values. In other words, in the case of assigning more balanced or fair weights to more important and less important terms (such that *TF1* does it relative to *TF2*), the effect of retrieval performance, or the MAP measure, for the final weightings remained almost stable while assuming that the most important term had a greater frequency count than that represented by an optimal alpha value.

The other important issue was that the final weighting *WTF1*, which was more balanced, had greater retrieval performance than *TF1*. However, performance gains of the presented weighting models were reasonable; the maximum retrieval performance (according to MAP) was below 10% percent of Okapi's BM25 weighting function.

On the strength of experimental inferences, in order to obtain an effective retrieval performance, we may say that *tf* within the document would be modeled (the TF component) on the basis of:

- Making the normalization of term frequencies regarding the variations of document lengths as accurate as possible, as is seen in the problem of z-score standardization (see section 4.2).
- Assigning balanced weights to the more important and the less important terms within a document, as possible.

In IR applications, the topic represented by a query is generally relevant to a subtopic of a document. Frequently, it is more valuable to weight the relevancy in terms of subtopics rather than the primary topic of a document. This means that the primary consideration of TF weighting must be focused on setting the balance between the subtopics and primary topic in order to calculate relevancy weighting. Therefore, it is possible to obtain good retrieval performance by a constructed TF component that is directly proportional to *tf*, if a balance can be established between less important lower frequencies and more important medium frequencies of terms and if a method for eliminating the portion of high frequencies that are functional terms (language-dependant words) is used. Our claim is also supported experimentally with the findings observed from the tests that were explained before.

Additionally, on the basis of our findings, the visibility of Luhn's claim in the TREC databases considered shows some variations. The degree of support seen in the databases varied from high to low in the order of TREC-8, TREC-6, and TREC-7. Meanwhile, the performance variation from high to low was seen in the same order of databases, which gives clue about the correlation between the degree of support of the claim and IR performance.

Future work is planned to investigate solutions to the relative weaknesses of intradocument *tf* modeling. Moreover, further works will focus on developing more efficient models based on Luhn's view. Some of such considered improvements are to construct alternative formulas that discriminate the estimation of term importance weights more fairly and to define alternative models, different from using z-scores, that are appropriate for Luhn's claim.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) within the scope of Project No. 107E192. The authors thank TÜBİTAK for supporting this project.

References

[1] M.E. Maron, J.L. Kuhns, "On relevance, probabilistic indexing and information retrieval", J. ACM, Vol. 25, pp. 216-244, 1960.

[2] H.P. Luhn, "A statistical approach to mechanized encoding and searching of literary information", IBM Journal Research and Development, Vol. 1, pp. 309-317, 1957.

[3] H.P. Luhn, "The automatic creation of literature abstracts", IBM Journal of Research and Development, Vol. 2, pp. 159-165, 1958.

[4] G. Salton, "Automatic text analysis", Science, Vol. 168, pp. 335-343, 1970.

[5] J. Minker, E. Peitola, G.A. Wilson, "Document retrieval experiments using cluster analysis", Journal of the American Society for Information Science, Vol. 24, pp. 246-260, 2007.

[6] S.E. Robertson, S. Walker, "Some simple approximations to 2-Poisson model for probabilistic weighted retrieval", in Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Dublin), New York, Springer-Verlag, pp. 232-241, 1994.

[7] K.S. Jones, S. Walker, S.E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments", Information Processing and Management, Vol. 36, pp. 779-840, 2000.

[8] G. Salton, A. Wong, C.T. Yu, "Automatic indexing using term discrimination and term precision measurements", Information Processing and Management, Vol. 12, pp. 43-51, 1976.

[9] G. Salton, C.S. Yang, "On the specification of term values in automatic indexing", Journal of Documentation., Vol. 29, pp. 351-372, 1973.

[10] K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, Vol. 28, pp. 11-21, 1972.

[11] S.P. Harter, "A probabilistic approach to automatic keyword indexing, Part I: On the distribution of specialty of words in a technical literature", Journal of the American Society for Information Science, Vol. 26, pp. 197-216, 1975.

[12] S.P. Harter, "A probabilistic approach to automatic keyword indexing, Part II: An algorithm for probabilistic indexing", Journal of the American Society for Information Science, Vol. 26, pp. 280-289, 1975.

[13] S.E. Robertson, K. Sparck Jones, "Relevance weighting of search terms", Journal of the American Society for Information Science, Vol. 27, pp. 129-146, 1976.

[14] W.S. Cooper, M.E. Maron, "Foundations of probabilistic and utility-theoretic indexing", Journal of the ACM, Vol. 26, pp. 67-80, 1978.

[15] W.B. Croft, D.J. Harper, "Using probabilistic models of document retrieval without relevance information", Journal of Documentation, Vol. 35, pp. 285-295, 1979.

[16] S.E. Robertson, C.J. van Rijsbergen, M. Porter, "Probabilistic models of indexing and searching", in Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval, Cambridge, England, pp. 35-56, 1980.

[17] N. Fuhr, "Models for retrieval with probabilistic indexing", Information Processing & Management, Vol. 25, pp. 55-72, 1989.

[18] H.R. Turtle, W.B. Croft, "A comparison of text retrieval models", The Computer Journal, Vol. 35, pp. 279-290, 1992.

[19] S.K.M. Wong, Y.Y. Yao, "On modeling information retrieval with probabilistic inference", ACM Transactions on Information Systems (TOIS), Vol. 13, pp. 38-68, 1995.

[20] J. Ponte, B. Croft, "A language modeling approach in information retrieval", in Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne), New York, ACM, pp. 275-281, 1998.

[21] D. Hiemstra, A.P. de Vries, "Relating the new language models of information retrieval to the traditional retrieval models", CTIT Technical Report TR-CTIT-00-09, Enschede, the Netherlands, Twente University, 2000.

[22] G. Amati, C.J. van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness", ACM Trans. Inf. Syst., Vol. 20, pp. 357-389, 2002.

[23] M. Porter, "An algorithm for suffix stripping", Program 14, pp. 130-137, 1980.

[24] S.E. Robertson, S. Walker, M. Beaulieu, "Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive", in the 7th Text Retrieval Conference NIST Special Publication 500:242, pp. 253-264, 1999.