International Journal of Natural and Engineering Sciences 5 (3): 31-35, 2011
ISSN: 1307-1149, E-ISSN: 2146-0086, www.nobel.gen.tr

Luhn's Point of View: Median-Based Term Weighting Schemes

İlker KOCABAŞ[1*] Bahar KARAOĞLAN[1] Bekir Taner DINÇER[2]

[1] International Computer Institute, Ege University, 35100, Bornova, Izmir, TURKEY

[2] Department of Computer Science, Muğla University, 48100, Kötekli, Muğla, TURKEY

***Corresponding Author**

e-mail:ilker.kocabas@ege.edu.tr

Received: June 27, 2011
Accepted: July 02, 2011

Abstract

In this study we replace the TF component of the TFxIDF term weighting method with a parameter derived from Luhn's claim on term importance. Luhn claims that the words with the mid frequencies are the most important ones, and the importance of a word fall as the frequency of the word increases or decreases. We take the median frequency of the words in a document as the base and assess the importance of a word by the difference between its frequency and the median frequency. The weighting functions are varied by two normalization approaches as using median itself and *standard deviation* of medians and tested on TREC-6 through TREC-8 adhoc tracks. The experimental results of the weightings using median itself, perform better retrieval than basic TFxIDF and BM25 with respect to MAP and R-P measures.

Key Words: Information retrieval, indexing, term importance

INTRODUCTION

Basically, the context of the text document is formed by the aggregation of the semantic of each term within the document itself; a term may be a letter, a number, a character or any combination of these that has a meaning. However, different terms within a document contribute different amounts to the semantic information conveyed. That is to say, the importance of each term which refers to its contribution measure may vary. If this variation of term importance can be reflected in representation of semantic information by means of weights given on index terms, the information content of a document can be characterized more precisely [1].

An approach to term weighting is 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 it is the first time of using term frequency notion for weighting terms, he presented a description of the relation between the term frequency (*tf*) within a text document and its informative content or significance in his later work [3]. His claim/description 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 in words by the following aspects:

(a) The terms with medium frequencies are more important than the terms that have low and 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 don't contribute significantly to the content of the text.

(b) Resolving power, the degree of strength in discriminating the content, of significant words in a text, reach peak at a point within the medium frequencies range (between the two cut-offs) and fall in both directions becoming almost negligible at the cut-off points.

Even though Luhn has put forth his claim just for 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 *the contribution of a term to the document context* is regarded as TF component of basic TF-IDF weighting scheme. Salton [4] and Minker et. al. [5] showed experimentally that using such a TF component on weighting index terms results superior retrieval performance over unweighted terms. Moreover, Robertson and his friends proposed alternative TF schemes [6,7]. Although these studies are inspired from the idea "taking *tf* into account when assigning weights to the terms", which was mentioned before by Luhn, they differentiate form Luhn's claim by the assumption: "all the time, *the weighting of a term with in a document* (TF) is directly proportional to *its frequency (tf)*". Briefly, this is the general assumption underlying the weighting methods that explicitly define 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 which are of high frequencies. However, the gap between the Luhn's description in degree of significance, as interpreted in case b above, and such approaches still exists.

One of the different weighting approaches based on combining the inter-document and intra-document term frequencies is the term discrimination value (TDV) [8]. In addition to Luhn's viewpoint, Salton et. al. [9] noted the impact of the frequency distribution of each term within the collection by examining the behaviors of TDV. Their conclusions are actually the expansion of Luhn's claim on collection-wide. On the other hand, TDV which is primarily intended to discriminate the vocabulary terms of a collection performs the same task that is aimed by inter document frequency (IDF) [10]. Hence, it is impossible to reach a judgment that TDV owns the same extensions as the Luhn's term's significance interpretation within the document boundary.

In our previous study [11], we investigated the validity of Luhn's point of view on term importance issue. Also in that work, *z*-scores were used for the quantitative expression of the claim described above. Our findings from broad experiments carried on several datasets support the validity of this claim but the performances of those weightings are not satisfactory for information retrieval.

In this study, we present several formulas based on Luhn's point of view for TF component in TFxIDF weighting schemes. Our endeavor is, in general, to determine the frequency of the most important term within a document as Luhn claimed, if it exists. In this study, we suppose that this frequency value is the median of the frequency of all words observed in a document. In the context of this consideration, we present several formulas for TF component in TFxIDF weighting schemes. We then tested their effectiveness in IR on TREC databases. In the following sections, our term weighting methods are explained precisely and experimental results are given.

MATERIALS AND METHODS

Median-Based TF×IDF Models

Any set of documents can be represented as *Term* × *Document* matrix *X*, shown in Figure 2. Rows and columns of matrix *X* represent t_i (*i=1..r*) terms and d_j (*j=1..c*) documents respectively. Each cell of *X*, x_{ij} indicates the number of occurrences of the term t_i in the document d_j. In the rest of paper, *word* and *term* are used interchangeably.

Suppose that all the terms seen in a document are sorted by their frequencies in either ascending or descending order, and the terms with the same frequency value are later grouped together. Thereafter, *the group of terms in the middle* (*median*) may be treated as the most important terms in accordance with Luhn's suggestion. The frequency of the median is therefore

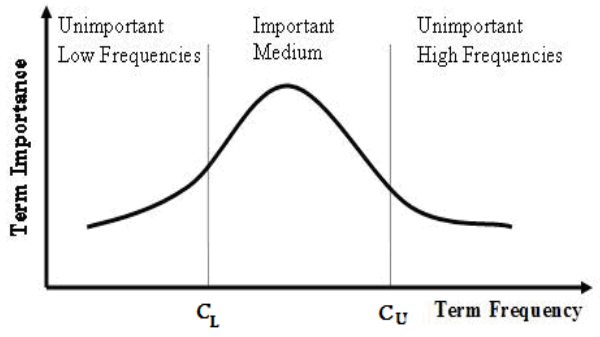

Figure 1. Relation between term importance and term frequency [3, adaptive]

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

Figure 2. Term × Document matrix

considered as the peak of the middle range frequencies in Figure 1. Thereby, we may quantitatively measure the importance of any term within a document in terms of the difference between its frequency and the median frequency of the document.

Definition: Each individual terms t_i in a document d_j belongs to a frequency class *f* with respect to its frequency x_{ij}. If all frequency classes observed for a document are sorted in ascending or descending order, then the *median* frequency (M_j) in d_j is the frequency class which is in the middle.

The main point of the median selection is to take the same frequency values observed with different terms into account only once.

Definition: US_{ij} is the measure of distance between the frequency x_{ij} of a term t_i in a document d_j and median M_j of terms in d_j.

Taking into account the Luhn's claim, we can further say that the importance of a term increases as US_{ij} decreases and it is at the peak when $US_{ij} = 0$, which indicates that the frequency of the term equals the median frequency, M_j in d_j. Thus, for any term t_i in a document d_j, *distance* US_{ij} and term importance (the weighting TF_{ij} of t_i) are inversely related to each other. This inverse relation is expressed as:

$$TF_{ij} \propto \frac{1}{US_j} \tag{1}$$

US_{ij} may simply be taken as the "absolute difference" between the term frequency and the median frequency, as given in Equation 2.

$$US_{ij} = \left| x_j - M_j \right| \tag{2}$$

For the purpose of avoiding the effect of the document length, US_{ij} may be normalized with the *standard deviation* of intra-document term frequencies as given in equation 3.

$$(US_{ij})_1 = \left| x_{ij} - M_j \right| \Big/ \sqrt{s_{mj}^2} \quad , \tag{3}$$

$$\sqrt{s_{mj}^2} = \sqrt{\frac{1}{r-1}\left(\sum_{i=1}^{r} x_{ij} - M_j \right)}$$

Consequently, the distance, US_{ij} of a term within a document becomes comparable with the ones of same term within other documents, regardless of its document length. The preliminary assumption under this sort of normalizations is

that the frequency distribution of a particular term over a set of documents may depend on some estimated features possessed by individual documents. However, what these features are not yet known at all. At this point, it seems more meaningful and accurate to include these features to the formulation *implicitly*, if possible, instead of some estimated ones. Here note that, the median frequency of a certain document should/might be affected by the same sort of features. In that sense, the document length dependency can be *indirectly* removed providing that the distance given in Equation 2 is defined in terms of median steps as given in Equation 4.

$$(US_{ij})_2 = \left| x_{ij} - M_j \right| \Big/ M_j \tag{4}$$

By using the two normalization methods given in Equation 3 (No:1), and Equation 4 (N0:2), TF_{ij} can be expressed quantitatively as two different functions that are shown in Equation 5 (a) and (b).

$$TF_{ij} = \begin{cases} \log_2\left(\dfrac{1}{US_j + 1} + 1\right) & \text{(a) TF1} \\ \log_2\left(\dfrac{1}{US_{ij}^{\,2} + 1} + 1\right) & \text{(b) TF2} \end{cases} \tag{5}$$

In these TF_{ij} functions, the value of US_{ij} distance is incremented by 1 in order to avoid the *division by zero* anomalies. Under the assumption of mutually exclusive behavior among terms, the ranking function should compute the degree of relevancy of each document with respect to a query as the multiplication of weights of the terms included. The other possible way to assure this assumption is to express the ranking function in additive form of the logarithmic transformations of actual term weights. Also by choosing base 2 for logarithm, functions generating values between 0 and 1 can be regarded as probability functions.

In our term weighting functions conforming to the TF×IDF schemes, Sparck Jones's *idf* [10] is used for IDF component; $idf = \log_2(N/n_o+1)$ where N is the total document numbers in a collection and n_o is the number of documents that a particular term t_i is observed. These functions are given in Equation 6, called as *TF1-IDF* and *TF2-IDF* in accordance to using *TF1* and *TF2* as TF component.

$$TF_j\, xIDF_i = \begin{cases} \log_2\left(\dfrac{1}{US_j + 1} + 1\right) \bullet idf & \text{(a) TF1 - IDF} \\ \log_2\left(\dfrac{1}{US_j^{\,2} + 1} + 1\right) \bullet idf & \text{(b) TF2 - IDF} \end{cases} \tag{6}$$

RESULTS

Experimental Setup

We carried out all of our experiments on TERRIER (TExt RetRIvER) 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) which is on disks 4 and 5. For this collection, we performed tests on each of 50 topics in TREC-6 ,TREC-7 and TREC-8. The TREC-6 test collection consists of about 2.1 GB data, of about 556000 documents, from the Congressional Record (CR), Financial Register (FR), Foreign Broadcast Information Service (FBIS), LA Times (LA) collections. In TREC-7 and TREC-8, the collection CR which includes large size documents was removed from indexing. After that, the average length of documents decreased to 512 tokens from 557 tokens.

Each of the topics has the same structure that consists of 3 fields. These are *title* which includes the most related words (one to three words), *description* which gives a wider explanation about query (one or two sentence), and a *narrative* which contains specific conditions on accepting or rejecting documents (a paragraph). In our experiments, we used the fields in two forms: in first case the query is composed of *only title* field (T-only), in second case the query is composed of *title and description* fields (TD).

In the indexing phase, Porter's stemming algorithm [12] 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 on the queries.

Experiments On Median-Based TF×IDF Schemes

Models constructed as median-based TF-IDF schemes in accordance with Luhn's point of view are explained in previous section. The weighting functions of such schemes are named as *TF1-IDF* and *TF2-IDF* with respect to TF components used. Also *TF1-IDF(α)* and *TF2-IDF(α)* represent those functions varying with respect to computation used for US_{ij}: $\alpha = 1$ for Equation 3 and $\alpha = 2$ for equation 4. The developed TF×IDF models run on TREC-6, TREC-7 and TREC-8 datasets for T-only and TD type queries.

Mean Average Precision (MAP), R-Precision, and precision values at 1, 5, 10, 30, 100 documents; which are the notations used P@1, P@5, P@10, P@30, P@100, respectively; are used as retrieval performance measures. We compared our models' performance results to basic TF×IDF scheme known as Robertson's TF [6] x Sparck Jones's IDF [10] and the ones of Okapi BM25 [13], broadly used weighting function.

Tests results of TREC-6 dataset are given in Table I. By using TF1-IDF(2) and TF2-IDF(2) weighting functions, approximately 5% more relevant documents (*RR*) are retrieved

Table 1. Performance Measures on Trec-6 Dataset (Query Type: T-Only)

Models	Performance Measures					
	RR	***MAP***	***R-P***	***P@5***	***P@10***	***P@100***
TF1-IDF(1)	2173	0.1708	0.2147	0.2880	0.2740	0.1508
TF2-IDF(1)	2161	0.1703	0.2123	0.2720	0.2700	0.1502
TF1-IDF(2)	2292	0.2046	0.2584	0.3720	0.3620	0.1700
TF2-IDF(2)	2300	0.2136	0.2639	0.4200	0.3700	0.1752
Basic TFxIDF	2156	0.2105	0.2544	0.4600	0.3960	0.1700
BM25	2173	0.2061	0.2545	0.4040	0.3740	0.1682

than using other functions. TF2-IDF(2) has the best performance measures on *MAP* (0.2136) and *R-P* (0.2639). Moreover, TF1-IDF(1) and TF2-IDF(1) retrieval perfomances are observed very poor, such that around %20 lower on MAP and R-P with respect to other retrieval performances. Basic TFxIDF performs better than others according to *P@5* and *P@10* measures.

Tests results of TREC-7 dataset are given in Table II. For all weighting functions, observed values of RR are nearly same. TF1-IDF(2) and TF2-IDF(2) obtain the best performance measures on MAP (0.1689 and 0.1708 respectively) and R-P (0.2239 and 0.2223 respectively). Moreover, TF1-IDF(1) and TF2-IDF(1) retrieval performances are observed very poor such that around %10 lower on MAP and R-P with respect to other retrieval performances. Basic TFxIDF and BM25 perform better than others according to P@5, P@10 and P@100 measures.

Tests results of TREC-8 dataset are given in Table III. By using TF1-IDF(2) and TF2-IDF(2) weighting functions, approximately 5% more relevant documents (RR) are retrieved than using Basic TFxIDF and BM25. Moreover the performance of these functions are higher nearly %10 than all others in terms of MAP, where TF2-IDF(2) has the best performance measures on MAP (0.2398) and R-P (0.2901). TF1-IDF(1) and TF2-IDF(1) retrieval performances are observed very poor, such that around %10 lower on MAP and R-P with respect to other retrieval performances. Basic TFxIDF performs better than others according to P@5 and P@10 measures.

Table 2. Performance Measures on Trec-7 dataset (Query type: T-only)

Models	Performance Measures					
	RR	***MAP***	***R-P***	***P@5***	***P@10***	***P@100***
TF1-IDF(1)	2134	0.1538	0.2125	0.3240	0.3060	0.1562
TF2-IDF(1)	2143	0.1519	0.2074	0.3200	0.3100	0.1576
TF1-IDF(2)	2161	0.1689	0.2239	0.3720	0.3440	0.1642
TF2-IDF(2)	2170	0.1708	0.2223	0.3640	0.3480	0.1672
Basic TFxIDF	2172	0.1632	0.2161	0.4160	0.3660	0.1686
BM25	2186	0.1641	0.2143	0.4200	0.3660	0.1692

Table 3. Performance Measures On trec-8 dataset (Query type: T-only)

MODELS	Performance Measures					
	RR	***MAP***	***R-P***	***P@5***	***P@10***	***P@100***
TF1-IDF(1)	2706	0.2203	0.2640	0.3880	0.3640	0.2016
TF2-IDF(1)	2743	0.2178	0.2649	0.3960	0.3700	0.2058
TF1-IDF(2)	2784	0.2341	0.2849	0.4120	0.3860	0.2164
TF2-IDF(2)	2812	0.2398	0.2901	0.4320	0.4180	0.2240
Basic TFxIDF	2670	0.2203	0.2804	0.4480	0.4240	0.2154
BM25	2672	0.2198	0.2780	0.4360	0.4220	0.2142

CONCLUSION

In this study, we formulated several TF weighting functions based on Luhn's claim on importance of words within a document. Also these TF functions are located in document ranking functions applicable to basic TF×IDF scheme. All of the experiments were carried out on TREC-6 through TREC-8 adhoc tracks. The experimental results of presented weightings based on median approach show that "using median itself rather than *standard deviation* of medians is more suitable for normalization". The weightings using median itself, TF1-IDF(2) and TF2-IDF(2) performs better retrieval than Basic TFxIDF and BM25 with respect to MAP and R-P measures. Especially, TF2-IDF(2) obtain 5-10% higher MAP performance than those at TREC-7 and TREC-8 datasets. On the other hand, BM25 and Basic TFxIDF show better retrieval performances with respect to P@5 and P@10 at all datasets.

Future work is planned to analyze the retrieval performances of presented weightings more deeply, such as experiments on other fields of topics, topic-based analysis, etc. Moreover further works will focus on developing more efficient alternative approaches in order to express the gap between term frequency and median of terms frequencies.

Acknowledgment

This work is supported by Scientific and Technical Research Council of Turkey (TUBITAK) within the scope of the project no: 107E192. The authors thank to TUBITAK for supporting this project.

REFERENCES

[1] Maron ME, Kuhns JL. 1960. On relevance, probabilistic indexing and information retrieval. Journal of ACM. 25(3):216-244.

[2] Luhn HP. 1957. A statistical approach to mechanized encoding and searching of literary information. IBM Journal Research and Development. 1(4):309-317.

[3] Luhn HP. 1958. The automatic creation of literature abstracts. IBM Journal of Research and Development. 2:159-165.

[4] Salton G. 1970. Automatic text analysis. Science. 168:335-343.

[5] Minker J, Peitola E., Wilson GA. 1973. Document retrieval experiments using cluster analysis. Journal of the American Society for Information Science. 24(4):246-260.

[6] Robertson SE, Walker S. 1994. Some simple approximations to 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the 7th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (Dublin), Springer-Verlag. N.Y. 232-241.

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

[8] Salton G, Wong A., Yu CT. 1976. Automatic indexing using term discrimination and term precision measurements. Information Processing and Management. 12(1):43-51.

[9] Salton G, Yang CS. 1973. On the specification of term values in automatic indexing. Journal of Documentation. 29(4):351-372.
[10] Jones KS. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 28(1):11-21.
[11] Kocabaş I, Dinçer BT, Karaoğlan B. 2011. Investigation of Luhn's claim on information retrieval. Turkish Journal of Electric Engineering and Computer Science (TJEECS). In press.
[12] Porter M. 1980. An algorithm for suffix stripping. Program 14. 130-137.
[13] Robertson SE, Walker S, Beaulieu M. 1999. Okapi at trec-7: Automatic adhoc, filtering, vlc and interactive. In the Seventh Text Retrieval Conference. NIST Special Publication 500:242. 253-264.