Journal of Intelligent Information Systems, 24:1, 61–85, 2005

A New Term Significance Weighting Approach

JIN ZHANG jzhang@csd.uwm.edu
University of Wisconsin-Milwaukee, Bolton Hall 532, P.O. Box 413, Milwaukee, Wisconsin 53201

TIEN N. NGUYEN tien@cs.uwm.edu
University of Wisconsin-Milwaukee, 3200 North Cramer Ave, EMS Building, Milwaukee, Wisconsin 53201

Received February 14, 2001; Revised April 26, 2004; Accepted April 28, 2004

Abstract. The authors present a new term significance measure that integrates term frequency retrieval characteristics, term frequency, document collection characteristics, and both the term depth and width distribution characteristics. A new concept, the term depth distribution, is introduced and its impact on the term significance is analyzed. The authors address the features of the new term significance measure from the angles of the impact of the variables (parameters) on it and the iso-significance contour analyses. An experimental study was conducted to compare the newly developed approach with two other popular approaches from the perspectives of both efficiency and effectiveness. The results show that the newly developed approach achieves satisfactory performance. Issues for further research on this topic are suggested.

Keywords: term significance, automatic term weighting, term weighting evaluation

1. Introduction

1.1. Related works

The objective of an information retrieval system is to provide its clients with satisfactory retrieval results. Toward this objective a retrieval result should be scientifically measured. Reasonable recall ratio and precision ratio of a retrieval result are two essential evaluation criteria for retrieval success. Achieving an acceptable recall ratio or precision ratio is a complicated and comprehensive process, influenced by numerous factors: quality of indexing, specificity and exhaustivity of indexing, user's information search expertise and experience, user's domain knowledge, database coverage, database organization structure, accuracy of users' need expression, search strategy formulation, an information retrieval system functionality and features, and so on. Among them quality of indexing is fundamental and extremely important. It is the first step of information retrieval process and the foundation of successful retrieval. The quality of indexing here refers to accuracy of selecting and weighting keywords for a document. Without high quality indexing, it is impossible to achieve a satisfactory search result. Determination of term importance plays a very important role in achieving high quality indexing. In addition, it is also the basis of automatic classification, automatic indexing, automatic abstracting, search feedback technique and a similarity measure (Debole and Sebastiani, 2003; Lai and Wu, 2002; Atlam et al., 2000;

Korfhage, 1997; Meadow, 1992; Rasmussen, 1992; Robertson et al., 1986; Salton, 1989; Umino, 1988; van Rijsbergen, 1979).

A wide variety of approaches have been addressed in weighting term importance. They range from the applicable to the theoretical, from the simple to the sophisticated. Some employ a genetic algorithm for assignment of weights to terms (Robertson and Willett, 1996), some use a scheme concept for weighting (Keen, 1991), some introduce a modeling method based on sources of documents to determine term importance (Wilbur, 1993), some borrow statistical theories to calculate term significance (Sparck Jones, 1972, 1973), some employ artificial neural network (Boger et al., 2001), some integrate the latent semantic technique in indexing (Gordon and Dumais, 1998), some apply probability theory to sovle the same problem (Greiff et al., 2002; Melucci, 1998; Ponte et al., 1998; Robertson et al., 1994; van Rijsbergen, 1977), and some just use a more practical and simple term frequency method (Greiff, 1998; Salton and Yang, 1973). Each approach has its disadvantages and advantages. Effects have been made to improve the existing weighting approaches (John, 2001; Zobel, 1998; Ro, 1988). In addition, a lot of research on comparisons among term weighting approaches (Jin et al., 2001; Salton and Buckley, 1988) has been done. Study on comparison between machine indexing and human indexing attempts to probe the nature of indexing (Anderson et al., 2001a, 2001b).

Clearly, term significance measure in a full text can be influenced by its frequency, type of a document (for example, scientific and technical paper, poetry, etc.), its context in the document, its function, its position in the document (for instance, in title, subtitle, abstract, introduction, conclusion and so on), and other factors. Among them, the context and function factors are related to a semantic environment that is extremely difficult to determine without fully understanding the full text while a term frequency can be easily calculated and has a close relationship to its importance. Therefore, term frequency is mostly applied to determine its significance in automatic information processing.

1.2. Motivation of research

Significance of a term within a document should not be calculated only based on the single document containing that term. It is widely recognized that documents in a database are not independent of each other. They affect each other in terms of their discriminative capacities in the database. For instance, if very few documents address a topic (subject) in a database, these documents are highly discriminative and they are easily distinguished from other documents from that topic. However, if many documents cover a same topic, it will decrease their discriminative capacities. Existing term weighting algorithms recognize and reflect the impact of document distribution in a database on term weighting.

Notice that multiple documents in a database can address a same topic but the degree to which they address the topic may vary. The extent to which they address the topic would also affect document discriminative capacity. Unfortunately, this factor is ignored in existing term weighting algorithms. The newly proposed term weighting algorithm attempts to add this new dimension to measurement of term significance and make the measurement more reasonable. It is the motivation of the research.

As one pioneer in this field, Luhn (1957) presented a simple measure depending on only raw term frequency to determine term significance: keep the high and discard the low. This method is simple and practical. It easily assures a high recall ratio due to retention of high frequency terms as indexing terms. However, the relationship between recall ratio and precision ratio indicates that the penalty of a high recall usually is a relatively low precision. A reasonable search result should maintain both at acceptable levels, that is, we should not emphasize one and ignore the other in retrieval process. In this case, Luhn's approach cannot meet the requirements for both reasonable recall and precision ratios.

As we know, a high precision ratio requires that indexing terms can strongly distinguish each document from others among a document collection. It has been recognized that term distribution in a document collection corresponds closely to this ability. The smaller the number of documents containing a term in a document collection, the better that term as a discriminator in the collection. Based on this idea, Salton and Yang (1973) came up with a new measure of term significance, the inverse document frequency measure. Supposing d_k, f_{ik} and N are defined as the number of documents containing term k, the raw frequency of term k in the document i, and the number of documents in a document collection respectively, term significance (or weight of term k in document i) w_{ik} is defined as:

$$w_{ik} = f_{ik} * \log\left(\frac{N}{d_k}\right) \tag{1}$$

Observe that in this equation the value of w_{ik} decreases as d_k increases and vice versa. Since it combines both the two factors—the distribution of a term within a certain document (term frequency f_{ik}) and its distribution in a document collection (logarithm of the ratio of the number of documents to the number of documents containing the term)—it was expected to get a nice result in both recall ratio and precision ratio in a search. Note that Eq. (1) does not factor in the length normalization, which addresses the impact of a document length on term significance. To ensure that all documents with different lengths have an equal chance of being retrieved, another version of Eq. (1) considering the length normalization was introduced by Salton et al. (1988, 1996).

$$w_{ik} = \frac{f_{ik} * \log(N/d_k)}{\sqrt{\sum_{j=1}^{m} (f_{ij})^2 * (\log(N/d_j))^2}} \tag{2}$$

Here m is the number of unique terms in a document vector space. Variable f_{ij} is defined as the frequency of term j in document i and d_j is defined as the number of documents containing term j. Without length normalization, the longer documents with more assigned terms and higher term frequencies would generate higher document similarities, and exhibit higher retrieval potential, than the shorter documents (Salton et al., 1996).

One noteworthy study of this topic was done by Sparck Jones et al. (1973). They discussed the logic of different types of term weighting approach and described experiments testing weighting schemes. The findings of their research showed that one type of weighting resulted in performance improvement. Let f_{ik} be the number of term (k) occurrences in document i, p_k the number of term (k) occurrences, and K the number of terms in the whole collection.

Then, term significance w_{ik} was defined as:

$$w_{ik} = f_{ik} * (K - \log(p_k)) \tag{3}$$

The method exploits term collection frequency p_k. The term frequency f_{ik} still plays the same role as that of the Salton equation.

Another examination of term significance was done earlier (Luhn, 1958). He recognized that very high-frequency terms in a document tended to have lower information-bearing value in terms of information retrieval, for example, "of", "with", "in", "the" and so on; and very low-frequency terms in a large document collection also had less significance.

Although the inverse document frequency measure, as one of the most popular measures, is widely accepted, it is not perfect yet. One of its weaknesses is that it simply employs a term frequency to multiply the logarithm of the ratio of the document numbers to the number of the documents containing the term. The frequency retrieval characteristics of the term significance mentioned by Luhn (1958) were not considered in the formula. Furthermore, the application of the logarithm of the ratio of the document numbers to the number of documents containing the term to indicate its ability of distinguishing it from others is not complete. In other words, it considers only the width characteristics of the term distribution, not the depth characteristics of the term distribution. The width and depth characteristics of a term distribution refer respectively to its distribution within the whole document collection (the number of documents containing the term) and its distribution within the documents containing the term (the number of the terms in these documents).

It is clear that neither of the algorithms (Sparck Jones et al., 1973; Salton et al., 1988) takes the term depth factor into consideration.

Figure 1 is a graphic display of the term depth distribution. D_i ($i = 1, 2, 3, 4, 5$ and 6) is a document containing a certain term. Increasing the number of documents containing a term will decrease its discriminative ability because when the term is used as a search term,

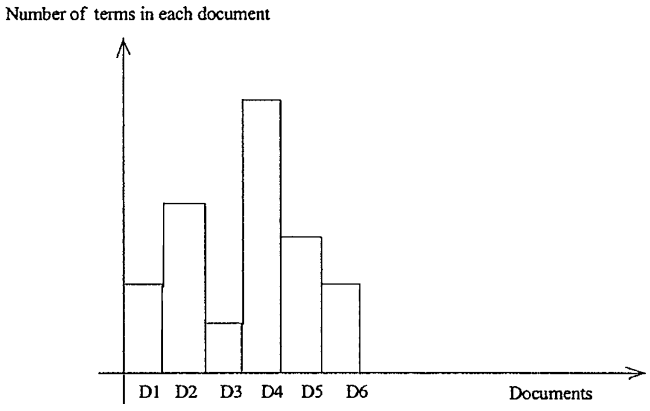

Figure 1. Term depth distribution.

more documents are retrieved (See figure 1, It is one-dimensional without the Y axis). As we know, it is the impact of the term width distribution on its significance.

If we consider a paragraph or sentence containing a term rather than a document as a basic retrieval object, the impact of the depth feature of the term on its significance can be recognized. For a full text document, not only the number of term occurrence in a document is recorded but also its position, e.g., paragraph number and/or sentence number is recorded. Increase of term frequency within the document will inevitably increase the probability that the document is retrieved from different paragraphs or sentences. It implies that if the number of a term in a document collection increases and the number of documents containing the term remains the same, the ability of the term as a discriminator decreases. It is the effect of term depth distribution on term significance.

Obviously, the depth characteristics of term distribution can affect term significance. To illustrate, say that two words A and B have the same width characteristics of term distribution: n different documents contain A and the same number of documents contain B; however, the depth characteristics of the two terms A and B are quite different: the frequencies of the term A within the n documents are much higher than those of the term B in the documents. That is, the number of the term A in the document collection is much higher than that of the term B in the n document collection. In this event, term significance of the terms A and B are definitely not the same even if the number of documents containing A is the same as that of documents containing B. Unfortunately, no of current significance measures can recognize this.

This paper proposes a new term significance measure to weight term importance more reasonably and more accurately by factoring in both term depth and term width characteristics.

In terms of the application of weighting term importance in information retrieval, there are three basic scenarios: one is to measure a document which contains a query term; the second is to select terms from a retrieved document to expand a query, a reverse process of the first; and the last is to choose some terms from a document as its surrogates. The presented model addresses the last situation.

2. Analysis and description of the new measure of weighting term significance

The above discussion shows that to reasonably, scientifically and accurately describe a term worthy of subject-indicating, the following dimensions should be taken into account:

(1) Term frequency in a document collection; (2) Frequency retrieval characteristics of term significance; (3) Document collection characteristics; and (4) Term distribution, including both its depth and width characteristics at a document collection level. In other words, the new term significance measure should integrate the four factors to solve the problem.

Equation (4) is the new measure:

$$W_{ik} = c^{-(f_{ik}-f_{ia})^2} * \log\left[\frac{N * D_k}{d_k * L_k}\right] \tag{4}$$

where f_{ia} is the middle value of frequency range in document i; f_{ik} is the raw frequency of term k in document i; L_k is the number of term k in the document collection; D_k is the number of all terms in documents containing term k; and W_{ik} is term significance of term k in document i, its weight. c (>0) is a constant used to adjust the impact of term frequencies on the weight. The other variables are defined as the same as Eq. (1).

We assume that d_k is not equal to zero, therefore, L_k is not equal to zero either; when d_k or L_k is equal to zero, the corresponding W_{ik} is defined as zero. In this way, the phenomenon in which W_{ik} becomes meaningless is avoided when d_k or L_k is equal to zero.

Luhn (1958) suggested that the terms located in the middle of a frequency range had a relatively stronger distinguishing ability than those located in the two ends of the frequency range. The first part of Eq. (4) describes this phenomenon.

$$S1 = c^{-(f_{ik}-f_{ia})^2} \tag{5}$$

Equation (5) manifests the effect of term frequency and frequency characteristics on term significance. We use a constant c to the power of $-(f_{ik} - f_{ia})^2$ rather than $1/(f_{ik} - f_{ia})^2$ to soften the effects of the variable changes on term significance. Another benefit of using a constant c is that by changing value of the constant c it allows users to control the degree to which the term frequency and frequency characteristics impact on term significance. In addition, this strategy can prevent the measure from meaningless when f_{ik} is equal to $f_{ia.}$ Equation (5) shows that terms in the middle of a frequency range have stronger impact on term significance.

$$S2 = \log\left[\frac{N * D_k}{d_k * L_k}\right] \tag{6}$$

In the second part $S2$, the ratio of N to d_k and the ratio of D_k to L_k fairly reflect the influences of the term width and depth distribution characteristics, respectively. The applications of parameters $\{N,\ D_k\}$ in Eq. (6) are the considerations of both the document collection characteristics and the characteristics of documents containing term k, respectively. The reason that we employ a logarithm of the ratio rather than a ratio is that this strategy can moderate the influence of variable changes on term significance. Equation (6) implies that the larger the number of documents containing a term, the less the impact of the term width distribution on the term significance, and vice versa. Equation (6) also suggests that the larger a term occurrence within documents containing it, the less the impact of the term depth distribution on term significance term, and vice versa; and the larger the number of all terms in documents containing the term, the stronger the impact of the term on its importance, and vice versa. The analyses show that the smaller the number of documents containing a term and its term occurrence in a document collection, the better that term as a discriminator.

The term significance model is based on term distribution characteristics. A term distribution can be divided into two levels: one is at whole document collection level (that is Eq. (6)) and another is at an individual document level (that is Eq. (5)). The frequency retrieval characteristics of a term within a document is one of the most important factors determining its significance at an individual document level while both its depth and width

characteristics indicate directly its distribution characteristics at whole document collection level. In other words, the two parts consist of a complete term distribution in a document collection. They are integrated and influence each other in terms of information retrieval.

In light of algorithm efficiency, application of exponential operation in the new algorithm may affect its efficiency. But the proposed algorithm is expected to outperform Salton's algorithm because the length normalization component of the Salton's algorithm (See the denominator in Eq. (2)) would slow down its processing.

A pseudo-code for the new algorithm is described as follows:

Repeat *each document i in a full-text-based database*
Do
    Parse its full text of document i;
    Use a keyword stop list to filter it;
    Calculate term k's frequency within a full text (f_{ik});
    Calculate f_{ia} the middle value of frequency range in document i;
Until *i reaches to N (the number of documents in the database)*
Establish a keyword list;
For *each term k in the keyword list*
Do
    Calculate L_k the number of term k in the database;
    Calculate D_k the number of all terms in documents containing term k;
    Calculate d_k the number of documents containing term k;
EndFor
For *each document i in the database*
    For *each term k in document i;*
    Calculate term weight W_{ik} based on Eq. (4);
    EndFor
EndFor

2.1. Analysis of the impacts of parameters and variables on term significance

Now we will address impacts of parameters and variables in the term significance measure on term significance w_{ik}.

2.1.1. Impact of f_{ik} and f_{ia} on term significance. In order to observe the impact of f_{ik} and f_{ia} on term significance, we isolate the values of other variables and parameters so that they can be isolated, where $N = 10000$, $L_k = 200$, $D_k = 1000$, $c = 1.2$ (Selection of c will be discussed in Section 2.1.4) and $d_k = 400$.

When f_{ik} is a variable, the value of f_{ia} is equal to 10 (for figure 2); when f_{ia} is a variable, the value of f_{ik} is equal to 20 (for figure 3). From Eq. (4), figures 2 and 3 are generated (See figures 2 and 3). The Y axis and X axis are significance and frequency of a term respectively in the both figures. Figure 2 tells us that when f_{ik} is equal to f_{ia}, term significance reaches its maximum value which depends on Eq. (4), when f_{ik} is getting far away

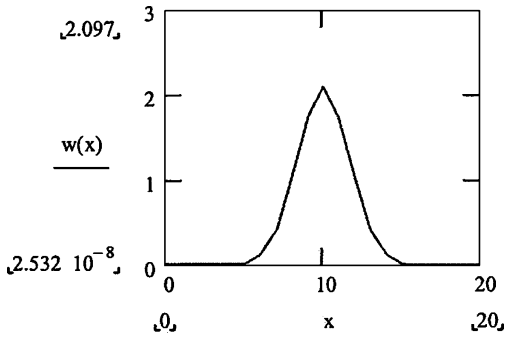

Figure 2. Impact of f_{ik}.

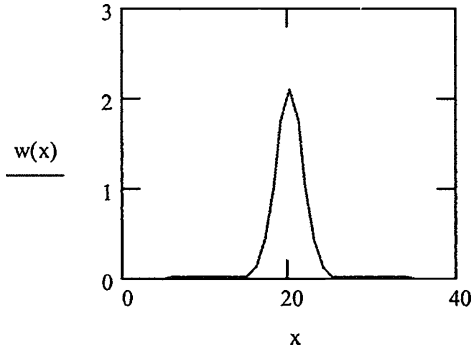

Figure 3. Impact of f_{ia}.

from f_{ia}, term significance w_{ik} will decrease. The curve in figure 2 exactly reflects Luhn's idea.

2.1.2. Impact of N and d_k on term significance. For the same reason mentioned above, we fix the following variables: let $f_{ik} = 20$, $f_{ia} = 15$, $D_k = 1000$, $c = 1.2$ and $L_k = 200$. When N is used as a variable, d_k is equal to 400 (see figure 4); when d_k is used as a variable, N is equal to10000 (see figure 5). The Y axis is term significance, the X axis is N in figure 4 and d_k in figure 5 respectively.

Figures 4 and 5 exhibit that when N and d_k increase respectively, the corresponding term significance w_{ik} will increase and decrease respectively.

2.1.3. Impact of L_k and D_k on term significance. Suppose $f_{ia} = 15$, $f_{ik} = 20$, $N = 10000$, $c = 1.2$ and $d_k = 400$. Figures 6 and 7 show their changes. The Y axis is term significance, X axis is L_k and D_k in figures 6 and 7 respectively. When L_k is a variable, D_k is equal to 1000; when D_k is a variable, L_k is equal to 200. Note that the curves in figures 4 and 5 are similar to those in figures 6 and 7 because the variables have similar positions in the newly developed measure.

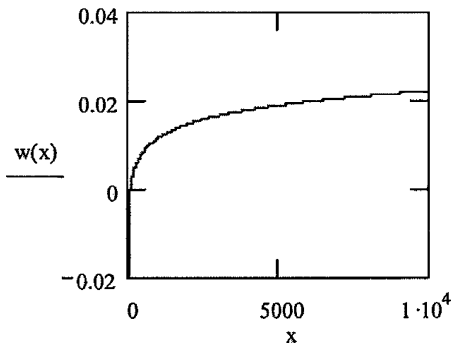

Figure 4. Impact of N.

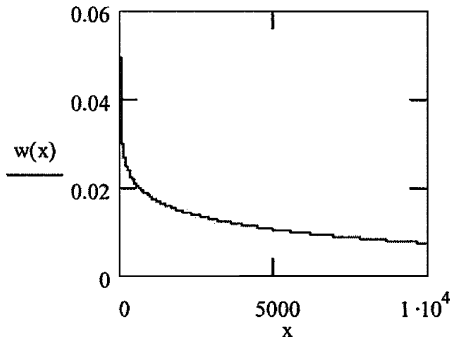

Figure 5. Impact of d_k.

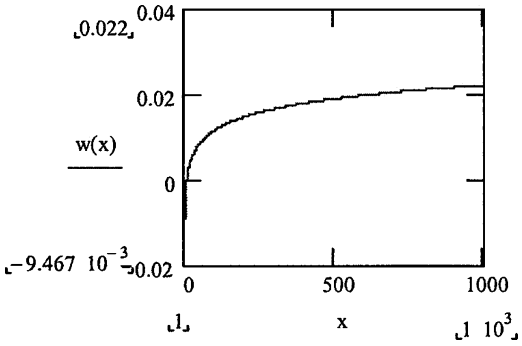

Figure 6. Impact of D_k.

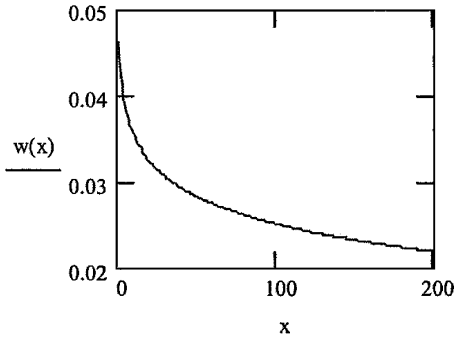

Figure 7. Impact of L_k.

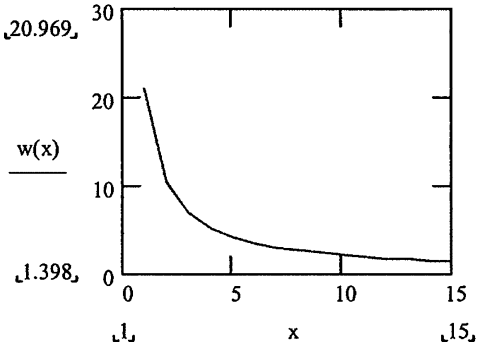

Figure 8. Impact of c.

2.1.4. Impact of c on term significance. Suppose $f_{ia} = 15$, $f_{ik} = 16$, $N = 10000$, $D_k = 1000$, $L_k = 200$, and $d_k = 400$. The Y axis is significance, the X axis is c in figure 8.

It is found that when the parameter c is set from 1 to 1.5, the impact on term significance is relatively stable. If c is selected in a certain range ($c < 1$, for instance), there will be no legitimate term significance value.

2.2. Iso-significance contour analyses

Iso-significance contour analysis is very important for understanding the new measure from different perspectives. It allows readers to investigate and observe relationships between two selected variables or parameters when term significance value remains at a constant level.

2.2.1. Iso-significance contour analysis based on f_{ik} and f_{ia}. From Eq. (4) we have the following new equation:

$$\log_c\left(\frac{\log\left[\frac{N*D_k}{d_k*L_k}\right]}{w_{ik}}\right) = (f_{ik} - f_{ia})^2$$

The variable f_{ia} has two solutions:

$$f_{ia} = f_{ik} + \left[\log_c \left[\frac{\log\left[\frac{N*D_k}{d_k*L_k}\right]}{w_{ik}} \right] \right]^{1/2} \tag{7}$$

$$f_{ia} = f_{ik} - \left[\log_c \left[\frac{\log\left[\frac{N*D_k}{d_k*L_k}\right]}{w_{ik}} \right] \right]^{1/2} \tag{8}$$

Suppose $N = 10000, D_k = 1000, L_k = 200, c = 1.2$ and $d_k = 400$, we have $w_{ik} = 1$ (for $f_{ia1}(x)$ *curves*); 0.001 (for $f_{ia2}(x)$ curves); and 0.000001 (for $f_{ia3}(x)$ curves) respectively. Figures 9 and 10 are generated by Eqs. (7) and (8) respectively.

The X axis and Y axis in figures 9 and 10 are term frequency and term significance respectively.

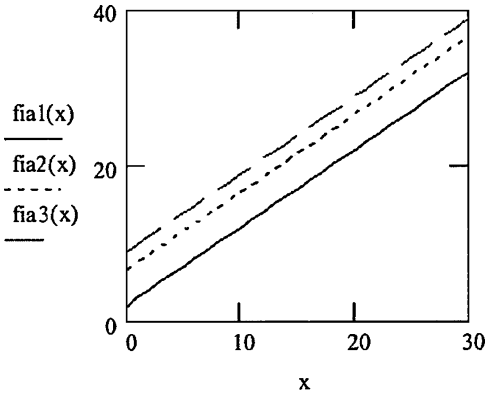

Figure 9. Iso-significance analysis of f_{ik} & f_{ia} (I).

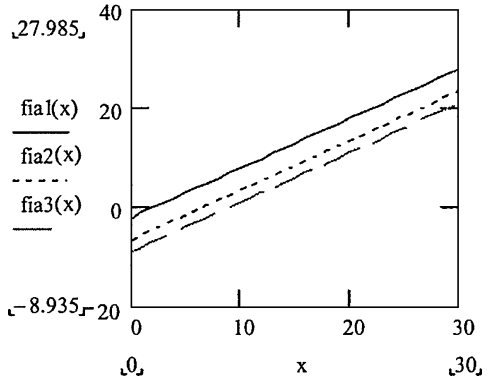

Figure 10. Iso-significance analysis of f_{ik} & f_{ia} (II).

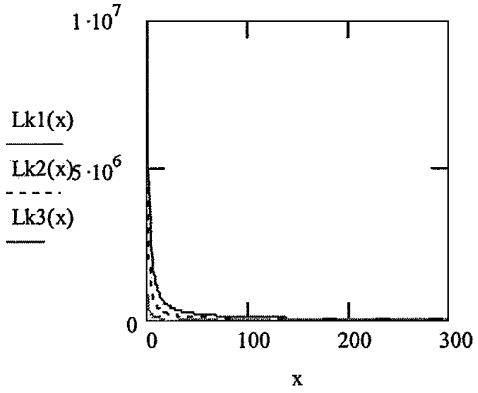

Figure 11. Iso-significance analysis of d_k & L_k.

Figures 9 and 10 show f_{ia} and f_{ik} have a linear relationship; the slopes of the straight lines are 1 regardless of the value of w_{ik}.

2.2.2. Iso-significance contour analysis based on L_k and d_k. From Eq. (4):

$$L_k = \frac{D_k}{d_k} * N * 10^{-w_{ik} * c^{[f_{ik} - f_{ia}]^2}} \tag{9}$$

Given $N = 10000$, $D_k = 1000$, $f_{ik} = 31$, $c = 1.2$ and $f_{ia} = 30$, we have $w_{ik} = 0.001$ (for $Lk1(x)$ curve); 0.3 (for $Lk2(x)$ curve); and 0.9 (for $Lk3(x)$ curve) respectively. See figure 11. The X axis is d_k and Y axis is L_k in figure 11. Figure 11 demonstrates L_k and d_k have a non-linear relationship. The smaller the w_{ik}, the higher the corresponding curve, and vice versa. It suggests that when d_k increases, L_k has to decrease dramatically to maintain a constant term significance value.

2.2.3. Iso-significance contour analysis based on f_{ik} and d_k. From Eq. (4):

$$d_k = \frac{D_k * N}{L_k} * 10^{-w_{ik} * c^{[f_{ik} - f_{ia}]^2}} \tag{10}$$

Suppose $N = 10000$, $L_k = 200$, $D_k = 1000$, $c = 1.2$ and $f_{ia} = 10$, we have $w_{ik} = 0.01$ (for $dk3(x)$ curve); 0.3 (for $dk2(x)$ curve); and 0.9 (for $dk1(x)$ curve) respectively. The X axis is f_{ik} and Y axis is d_k in figure 12. From figure 12, we find that the smaller the value of w_{ik}, the higher the corresponding curve. Note that the curves are symmetric against $X = f_{ia}$. It means that when term k appears frequently in documents, the number of documents containing k should be greater in order for the term to be significant, while when k is not frequent or is very frequent—it may appear in a few terms to be significant.

2.2.4. Iso-significance contour analysis based on f_{ik} and L_k. The required figure can be produced based on Eq. (9). Suppose $N = 10000$, $d_k = 400$, $D_k = 1000$, $c = 1.2$ and

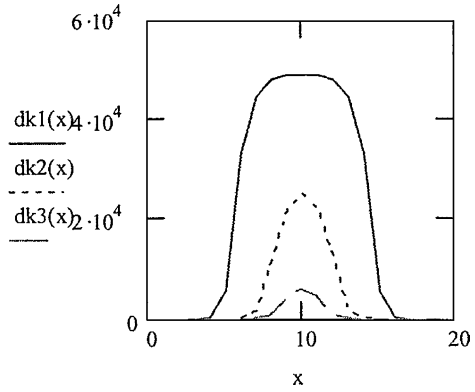

Figure 12. Iso-significance analysis of f_{ik} and d_k.

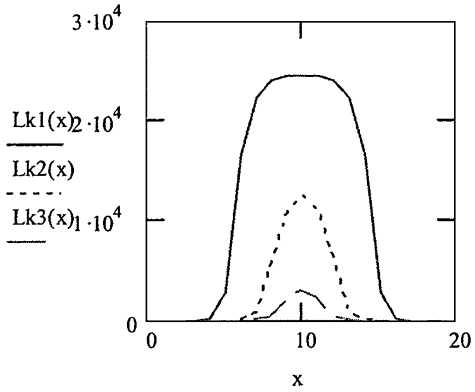

Figure 13. Iso-significance analysis of f_{ik} & L_k.

$f_{ia} = 10$, we have $w_{ik} = 0.01$ (for $Lk1(x)$ curve); 0.3 (for $Lk2(x)$ curve); and 0.9 (for $Lk3(x)$ curve), respectively.

The Y axis is L_k and X axis is f_{ik} in figure 13. Obviously, the smaller the value of w_{ik}, the higher the corresponding curve and vice versa. The curves are similar to those in figure 12. It implies that when f_{ik} is around a medium value, L_k should be great in order for the term to be significant, while f_{ik} is not frequent or is very frequent—it may appear in a few terms to be significant.

2.2.5. Iso-significance contour analysis based on N and D_k. We can get the following equation directly from Eq. (10) to produce required curves.

$$N = \frac{d_k * L_k}{D_k} * 10^{w_{ik} * c^{[f_{ik} - f_{ia}]^2}} \tag{11}$$

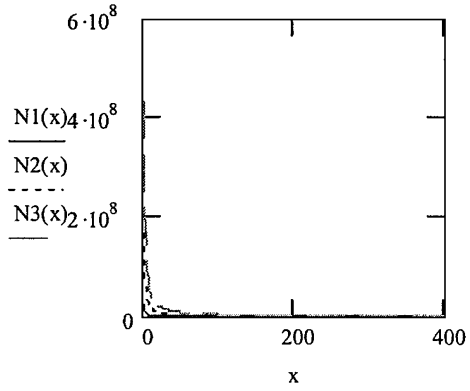

Figure 14. Iso-significance analysis of N and D_k.

Suppose $d_k = 400$, $L_k = 200$, $f_{ik} = 12$, $c = 1.2$ and $f_{ia} = 10$, we have $w_{ik} = 1.2$ (for $N1(x)$ curve); 1.6 (for $N2(x)$ curve); and 1.8 (for$N3(x)$ curve) respectively. See figure 14. The contour curves show the smaller the value of w_{ik}, the lower the corresponding curve and vice versa. The X axis is D_k and the Y axis is N in figure 14. It means that that when D_k increases, N decreases dramatically to maintain a constant term significance value.

2.2.6. Iso-significance contour analysis based on the N and f_{ik}. We can directly use Eq. (11) to produce the curves. Suppose $D_k = 1000$, $L_k = 200$, $d_k = 400$, $c = 1.2$ and $f_{ia} = 3$, we have $w_{ik} = 0.2$ (for $N1(x)$ curve); 0.21 (for $N2(x)$ curve); and 0.225 (for $N3(x)$ curve) respectively. The curves show the smaller the value of w_{ik}, the lower the corresponding curve and vice versa. See figure 15. The X axis is f_{ik} and the Y axis is N in figure 15. It is clear that the database contains many documents, a term provides the same contribution/significance when its frequency in a document is very low or very high.

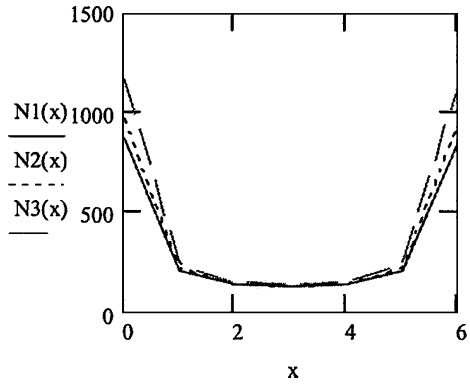

Figure 15. Iso-significance analysis of N & f_{ik}.

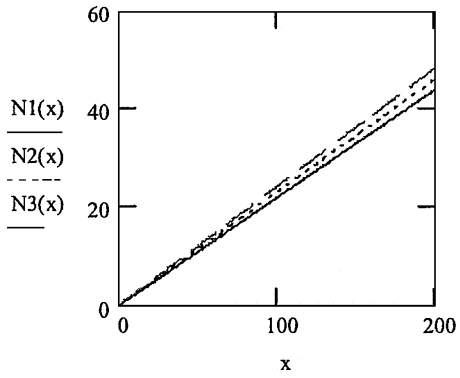

Figure 16. Iso-significance analysis of N and d_k.

2.2.7. Iso-significance contour analysis based on N and d_k. We can directly use Eq. (11) to produce the curves.

Suppose $L_k = 200$, $D_k = 1000$, $f_{ia} = 30$, $c = 1.2$ and $f_{ik} = 32$, we have $w_{ik} = 0.02$ (for $N1(x)$ curve); 0.03 (for $N2(x)$ curve); and 0.04 (for $N3(x)$ curve) respectively. The curves show the smaller the value of w_{ik}, the lower the corresponding curve and vice versa. All curves are linear and start from the origin in figure 16. The X axis is d_k and the Y axis is N in figure 16.

2.2.8. Iso-significance contour analysis based on D_k and L_k. We use the following equation to generate the figure.

$$D_k = \frac{d_k{}^* L_k}{N} * 10^{w_{ik} * c^{[f_{ik} - f_{ia}]^2}} \tag{12}$$

Suppose $N = 10000$, $d_k = 400$, $f_{ia} = 30$, $c = 1.2$ and $f_{ik} = 32$, we have $w_{ik} = 0.2$ (for $Dk1(x)$ curve); 0.3 (for $Dk2(x)$ curve); and 0.4 (for $Dk3(x)$ curve) respectively (See figure 17). They are three linear lines with different slopes. The X axis and Y axis are L_k and D_k respectively.

2.2.9. Iso-significance contour analysis based on D_k and f_{ik}. Equation (12) can be used to yield the curves. See figure 18. Suppose $N = 10000$, $L_k = 200$, $d_k = 400$, $c = 1.2$ and $f_{ia} = 3$, we have $w_{ik} = 0.2$ (for $Dk1(x)$ curve); 0.21 (for $Dk2(x)$ curve); and 0.225 (for $Dk3(x)$ curve) respectively. The X axis and Y axis are f_{ik} and D_k respectively.

2.2.10. Iso-significance contour analysis based on D_k and d_k. Equation (12) can be used to yield the contour. See figure 19. Suppose $N = 10000$, $L_k = 200$, $f_{ia} = 30$, $c = 1.2$ and $f_{ik} = 32$, we have $w_{ik} = 0.2$ (for $Dk1(x)$ curve); 0.3 (for $Dk2(x)$ curve); and 0.4 (for $Dk3(x)$ curve) respectively. The X axis and Y axis are d_k and D_k respectively. They are three straight lines starting from the origin.

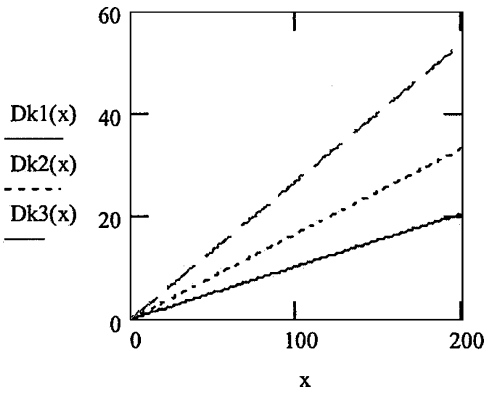

Figure 17. Iso-significance analysis of D_k and L_k.

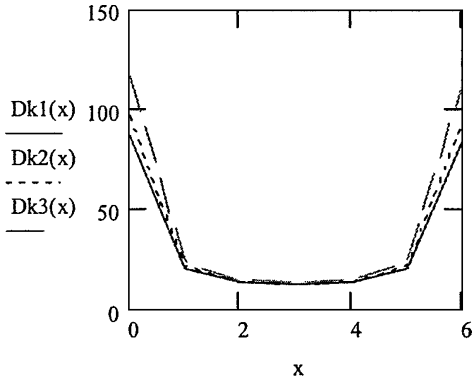

Figure 18. Iso-significance analysis of D_k and f_{ik}.

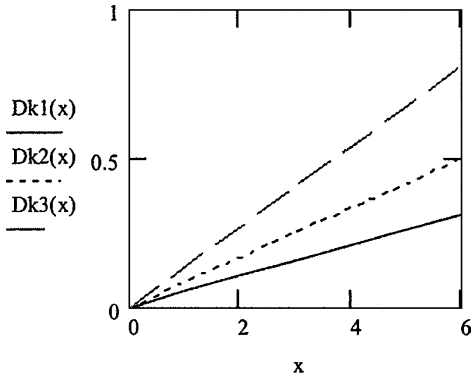

Figure 19. Iso-significance analysis of D_k and d_k.

3. Experimental study

An experimental study was conducted to examine and evaluate performance of the newly introduced method. Performance of the newly developed approach was compared with two popular term weighting approaches. One of them is the popular method "Inverse Document Frequency" proposed by Salton (1988). Note that the formula of "Inverse Document Frequency" used in this experimental study is not the original version (Salton and Yang, 1973), but the length-normalized version (Salton et al., 1988). The authors concluded that the length normalization version was much better than the previous one because it took into consideration of document length, which addresses the impact of document length on term significance. The normalized term weighting approach is shown in Eq. (2).

Another method (Sparck Jones, 1973) was also compared with the newly developed method. This proposed formula is shown in Eq. (3).

Theoretical analyses of the newly developed approach have been addressed in previous sections, but its soundness also requires experimental evidences to support it. The aim of this experimental study was to investigate whether the new term significance weighting method achieved a better performance than either the Salton's method or Sparck Jones' method from two perspectives: effectiveness and efficiency.

3.1. Statement of the hypothesis

To examine effectiveness of the newly developed method, we investigated performance of this method against both the Salton's method and Sparck Jones's method respectively. Proposed hypotheses are:

H1: The newly developed algorithm achieves better performance than that of the Salton's algorithm.

H2: The newly developed algorithm achieves better performance than that of the Sparck Jones' algorithm.

3.2. Methodology

In order to compare effectiveness of the three term weighting algorithms, measurement for keyword extraction accuracy must be clearly defined to examine performance. The three algorithms were used to extract the same number of keywords from each document. Keywords extracted from different algorithms may vary even they come from the same document. It is the differences that lead to different performance of the three algorithms. These keywords from different algorithms were collected and associated with the document where they were extracted. After subjects read content of a document, they will make keyword judgment for the document. That is, they were required to select most relevant keywords from these keywords based on the content of that document. The selected keywords were regarded as most relevant keywords for the document and were used to calculate the key work extraction accuracy rate for each term weighting algorithm. The keyword extraction accuracy rate for

a term weighting algorithm is defined as the ratio of the number of overlapping keywords between selected keywords by subjects and extracted keywords from the algorithm vs. the number of selected keywords by subjects. It is clear that in this case, the number of the selected keywords by subjects is the same for the three algorithms, and the number of overlapping keywords between selected keywords by subjects and extracted keywords for each algorithm varies. It is obvious that the larger a value of keyword extraction accuracy rate for an algorithm, the better performance of the algorithm.

The method was applied to all documents and keyword extraction accuracy rates for the three algorithms from all documents were collected. Then performance between the newly proposed algorithm and Salton' algorithm is compared and performance between the newly proposed algorithm and Spark Jones' algorithm is compared based on their keyword extraction accuracy rates via a T-test, respectively.

Here q automatic indexing algorithms $(M_1, M_2, \ldots, M_q)$ are compared. An extracted term set for a specific document D_i is produced based on the following procedure. Each algorithm is applied to extract a set of keywords for a document. Method M_1 generates a set of terms $R_{1i} = \{k_{11}, k_{12}, \ldots, k_{1n}\}$ for document D_i. Method M_2 produces $R_{2i} = \{k_{21}, k_{22}, \ldots, k_{2m}\}$ for document $D_i, \ldots$ etc. Then make a union for all R_{ji}'$s(j = 1, \ldots, q)$ to yield a new set: the *reference set* for document D_i. Finally, subjects make an ultimate judgment on final relevant keywords for that document based on the *reference set*, forming a new *standard set* $S_i = \{h_{1i}, \ldots, h_{ti}\}$ for document D_i. Notice that the *standard set* for a document is a subset of the *reference set*.

Observe that in figure 8, when a value of the constant c is between 1 and 1.5, its impact on term weight w is relatively stable. Therefore, in this experimental study, c is set to 1.2.

The keyword extraction accuracy rate, used to evaluate the performance of a weighting approach M_k $(k = 1, \ldots, q)$ for document D_i, is defined as

$$E_{ki} = \frac{|R_{ki} \cap S_i|}{|S_i|} \tag{13}$$

where $|S|$ refers to a size of the set S. So, document D_i, will correspond to an E_{ki}. It means that E_{ki} is the ratio of the relevant keywords judged by subjects for document D_i against a specified weighting approach M_k to all relevant keywords judged by subjects for document D_i against all automatic weighting approaches. In Eq. (13), $R_{ki} \cap S_i \subseteq S_i$ and in most cases, $R_{ki} \cap S_i \subset S_i$.

The participants were 20 students in computer science, library & information science, and business at the University of Wisconsin-Milwaukee. The subjects had backgrounds in both natural sciences and social sciences. The database used in this study is an *Associated Press* (*AP*) database containing more than 600 news reports in 1989. The database comes from TREC. Fields in a record include document number, title, sub-headline, author, date, and full text, but only full text and title were presented to subjects. An application was developed which extracted keywords from a full text based on the three automatic term weighting methods. The program was written in *Microsoft Visual C++ 6.0*, running on *Windows 1998/2000* and *NT*.

Experimental procedure is addressed as follows.

(1) *Preprocessing data*: The program was developed to parse a full text and collect data. Three distinct databases for the three term weighting methods were built for later keyword relevance judgment analysis.
(2) *Determination of the reference set for each document*: In this experimental study, five keywords were extracted for each individual document. Each weighting approach generated a keyword list. Each keyword in the keyword list was associated with a weight value. The keywords were ranked by their weight values. Five keywords with the highest weight scores were kept as indexing keywords for that document. It implies that there is not a constant weight threshold used to select keywords across all documents. Each document might have a different weight threshold for selection and therefore it was dynamic.

    The reason that only 5 keywords were extracted from each document is that the database used in this experimental study is news-oriented and topic of a document is relatively specific, 5 terms are enough to cover it. When contents of the database change and the number of extracted terms from a document increases, it may yield different results. According to the algorithm, potential weight for a term with a very low frequency (For instance, 1) is very small. Therefore, these terms with very low frequencies are filtered.

    After 5 keywords were extracted using a weighting approach, each approach had 5 keywords for a document. Finally, a union operation was performed on all keywords from the three different weighting approaches to form a reference set for that document. In most cases, the size of the reference set was larger than 5 if keywords for that document from three different weighting approaches were not totally overlapped. However, if the size was equal to 5, it indicated that the 5 keywords extracted from the three weighting approaches were exactly the same. In this case, an additional procedure followed up: increasing the number of extracted keywords (>5) for each weighting method until the total keyword overlapping was avoided.
(3) *Determination of the standard set for each document*: Each document consisted of a title, its full text, and the reference set of keywords. These were provided to each subject. Each subject was also required to select the 5 most relevant terms from the reference set after reading all information related to that document. These 5 keywords would be a standard set for that document.
(4) *Measurement of performance*: After a standard set for each document was determined, a value of E_{ki} for a weighting approach was computed based on Eq. (13).
(5) *Calculated E_{ki} for each of all documents*: Kept doing steps (1) to (4) until all documents were processed.

3.3. Data analysis

After experimental preparation, design and data collection, collected data were used to examine the proposed term weighting methods. The statistical software package *MINITAB*

Table 1. Comparison between the new algorithm and Salton's algorithm.

Two Sample T-Test and Confidence Interval

```
Two sample T for Zhang vs Salton

          N      Mean     StDev   SE Mean
zhang   100     0.826     0.131     0.013
salton  100     0.750     0.143     0.014

95% CI for mu Zhang - mu Salton: ( 0.038,  0.114)
T-Test mu zhang = mu salton (vs >): T = 3.90  P = 0.0001  DF = 198
Both use Pooled StDev = 0.137
```

was used in this phase to compare performance of the three methods. Term significance level α in this experimental study was 0.05.

There are two ways to judge a test result: the p-value approach and test statistic approach. The former bases acceptance of a hypothesis on a condition p-value $> \alpha$ at the $100.\alpha\%$ significance level if the comparison between two factors is set to be "equal to" or a condition p-value $\leq \alpha$ at the $100.\alpha\%$ significance level if the comparison between two factors is set to be "greater than". The latter approach makes the decision whether a hypothesis is rejected or accepted on a critical value at a significance level α.

Now let us analyze and examine the hypotheses $H1$ and $H2$.

H1. The new algorithm achieves better performance than that of the Salton's algorithm.

The result shows that the hypothesis $H1$ is accepted since the statistic t-test is set as ">" and p-value (0.0001) $< \alpha$ (0.05). Table 1 shows that the mean of the new algorithm (0.826) is larger than that of the Salton's algorithm. The smaller deviation of the new algorithm (0.131) than that of the Salton's algorithm (0.143) shows that performance of the former is more stable than the latter.

H2. The new algorithm achieves better performance than that of the Sparck Jones' algorithm.

The result of the hypothesis $H2$ is rejected since p-value (0.49) $> \alpha$ (0.05). Table 2 shows that the mean of the new algorithm (0.826) is the same as that of the Spark Jones's algorithm. The smaller deviation of the Spark Jones's algorithm (0.128) than that of the new algorithm (0.131) illustrates performance of the former is more stable than the latter.

The analytic results demonstrate that the newly developed term significance weighting method achieves better effectiveness performance than that by Salton and there is no significant difference between the new method and the Sparck Jones method.

Table 2. Comparison between the new algorithm and the Sparck Jones' algorithm.

Two Sample T-Test and Confidence Interval

```
Two sample T for Zhang vs Sparck Jones

                N      Mean     StDev   SE Mean
zhang         100     0.826     0.131     0.013
sparck jones  100     0.826     0.128     0.013

95% CI for mu zhang - mu sparck jones: ( -0.036,  0.037)
T-Test mu zhang = mu sparck jones (vs >): T = 0.03 P = 0.49
DF = 198
Both use Pooled StDev = 0.130
```

3.4. Efficiency analysis

In this experimental study, efficiency performance of the three methods was also examined. To compare efficiency of the new method to the others, the number of extracted keywords per second was defined to measure computational complexity for each approach. Final results in this category were provided in Table 3.

The three algorithms were applied to the same database. First all documents were processed and terms from each of documents were extracted. Then a term characteristics table was established. For each term in the table, it includes its term frequency in a document, its term frequency in a database, the number of documents containing that term and other related information. Finally, each of the three algorithms was used to calculate term weight based on the term characteristics table. Processing time for each algorithm and the number of extracted terms for each algorithm were recorded.

The number of documents is 636, the number of words in the document collection is 300,434, and the number of extracted terms is 19,238.

It is apparent that the Sparck Jones approach is the best among the three approaches in terms of efficiency. The newly developed approach is better than the Salton approach with respect to computational complexity. In terms of computational complexity, it is not surprising that the *Spark-Jones* algorithm achieves the best performance because it involves only simple multiplication operation and logarithmic operation (See Eq. (3)). The newly proposed algorithm has both logarithmic operation and relatively complex exponential operation (See Eq. (4)). The Salton's algorithm includes not only multiplication operation and logarithmic operation but also a very complex dominator (See Eq. (2)). Fortunately, algorithm efficiency is not top priority consideration due to the fact that basically a term weigh algorithm is used only in database construction not real time query response.

Table 3. Summary of statistical information for efficiency analysis.

Methods	Keywords per second
The new method	9.16
The Sparck Jones method	12.83
The Salton method (with length normalization)	4.28

4. Conclusion

As one fundamental area in the information retrieval field, a term significance measure has both theoretical and practical importance. With development of automatic feedback technique and the Internet search engines, a term significance measure would become more and more important. The discriminative ability of a term plays an extremely important role in weighting its significance. Identifying and analyzing the factors which affect the discriminative ability of a term in a document collection is crucial for the development of a new term significance measure. The newly developed term significance measure tries to define term significance from four different dimensions (term frequency, term frequency characteristics, document collection characteristics, and term depth and width distribution characteristics) and to integrate them into term weighting. Six different variables and parameters are involved in the newly developed term significance measure.

Term frequency, average term frequency and their combination reflect the frequency and the frequency characteristics; the application of a document collection size is the consideration of whole document collection characteristics; ratio of a document collection size to the number of documents with a certain term illustrates the term width distribution characteristics; and ratio of the number of all terms in documents with that term to the number of the term in a document collection presents the term depth distribution characteristics in term significance.

Analysis of each variable or parameter in the term significance measure gives users a clear picture about term significance, mutual impacts and relationships among variables or parameters, and helps them to appropriately apply this term significance measure.

Iso-significance contour analyses allow readers to further understand features of a term significance measure from a quite different angle. It provides users with rich information about interaction between two selected variables or parameters and their effects on term significance. These discussions are extremely important for a user to manipulate them correctly.

Study shows that there are some similarities between the impacts of D_k and N as well as L_k and d_k on the term significance measure. The variable pairs or parameter pairs, such as f_{ia} and f_{ik}, N and d_k, as well as D_k and L_k, have relatively close relationships in terms of their natures and their influences on the term significance measure. Based on this idea, we can divide them into three different groups in iso-significance contour analysis: $\{f_{ia}, f_{ik}\}$, $\{N, d_t\}$, and $\{L_k, D_k\}$, where f_{ik}, d_t and L_k are key variables in each group.

The application of the constant c is to moderate the impact of term frequency in the method on term weight. It is found that when the constant $c$ranges from 1.0 to 1.5, the impact is relatively stable.

This study introduces not only a new algorithm for term weighting but also an evaluation method for a term weighting algorithm. An experimental study was conducted to examine its performance against two other weighting approaches. The findings demonstrated that the new term significance weighting method achieves better performance than the Salton's approach and the same performance as that of the Sparck Jones' approach in terms of effectiveness. It achieves better efficiency performance than the Salton's approach.

It is apparent that the new algorithm is not better in terms of algorithm effectiveness and it is less efficient than Spark-Jones' algorithm. Notice that this experimental study was conducted in a small collection. From Eq. (3), we know that the relationship between parameter K (the number of terms in the whole collection) and term weight is linear the Spark-Jones' algorithm. That implies that if this algorithm is applied to a small size data collection (a small K), either term frequency f_{ik} and term collection frequency p_k in Eq. (3) still plays an appropriate role in term weight. However, once the Spark-Jones' algorithm is applied to a very large collection, K will be extremely large compared to the other two parameters (f_{ik} and p_k). In that case, the impact of K in Eq. (3) will be overwhelmingly dominant and it will significantly overshadow the impact of either f_{ik} or p_k on term weight. Therefore, their impacts would be ignored in the equation and this may reduce its algorithm effectiveness greatly.

On the other hand, due to the simplicity of the Spark-Jones' algorithm, it outperforms both the newly proposed algorithm and Salton's algorithm.

From Eq. (4), the constant c plays an important role in the new algorithm. Change of c value can be used to adjust the impact degree to which the term frequency and frequency characteristics impact on term weight. In this study, only one c value was used to calculate term weights for extracted terms. It is clear that if a different c is selected, the algorithm would weight these terms differently. If a group of c values rather than a single c were employed in the study, a wider range of comparisons among the three algorithms would have been made.

Notice that if $c^{-(f_{ik}-f_{ia})^2}$ is replaced by $1/(f_{ik} - f_{ia})^2$ (Or $1/|f_{ik} - f_{ia}|$) (Note: when $f_{ik} = f_{ia}$, then$1/(f_{ik} - f_{ia})^2$ (Or $1/|f_{ik} - f_{ia}|$) is defined as 1) in Eq. (4), the term frequency and frequency characteristics can still be preserved. And the replacement would improve its efficiency significantly because exponential operation is removed and the computational complexity is reduced.

It would be premature to draw general conclusions regarding the newly developed approach on the basis of this study because the database used consisted of a relatively small number of records of news-related full-text documents, and the number of subjects was not very large. Performance of the approaches may vary with the size and coverage of the database, or subject size. For example, as the number of documents increases, it may have a strong negative impact on term weighting effectiveness like the Spark Jones' algorithm. In this sense, application of the algorithms to a large data set can be one of future research directions.

Other further research on this term significance measure includes: control of the combination of parameters in the formula; iso-significance contour analysis based on multiple variables or parameters (more than two); and its potential applications in automatic information processing like abstracting and indexing.

References

Anderson, J.D. and Perez Carballo, J. (2001a). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part I: Research, and the Nature of Human Indexing. *Information Processing and Management*, 37(2), 231–254.

Anderson, J.D. and Perez Carballo, J. (2001b). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part II: Machine Indexing, and the Allocation of Human Versus Machine Effort. *Information Processing and Management*, 37(2), 255–277.

Atlam, E.S., Fuketa, M., and Morita, K. (2000). Similarity Measurement Using Term Negative Weight and Its Application to Word Similarity. *Information Processing and Management*, 36(5), 717–736.

Boger, Z., Kuflik, T., and Shoval, P. (2001). Automatic Keyword Identification by Artificial Neural Networks Compared to Manual Identification by Users of Filtering Systems. *Information Processing and Management*, 37(2), 187–198.

Debole, F. and Sebastiani, F. (2003). Information Access and Retrieval: Supervised Term Weighting for Automated Text Categorization. In *Proceedings of the 2003 ACM Symposium on Applied Computing* (pp. 784–788). Melbourne, Florida: ACM.

Gordon, M.D. and Dumais, S. (1998). Using Latent Semantic Indexing for Literature Based Discovery. *Journal of the American Society for Information Science*, 49(8), 674–685.

Greiff, W.R. (1998). A Theory of Term Weighting Based on Exploratory Data Analysis. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11–19). Melbourne, Australia: ACM.

Greiff, W.R., Morgan, W.T., and Ponte, J.M. (2002). Information Retrieval Models: The Role of Variance in Term Weighting for Probabilistic Information Retrieval. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 252–259). New York, NY: ACM.

Jin, R., Falusos, C., and Hauptmann, A.G. (2001). Meta-Scoring: Automatically Evaluating Term Weighting Schemes in IR Without Precision-Recall. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 83–89). New Orleans, Louisiana: ACM.

John, W.W. (2001). Global Term Weights for Document Retrieval Learned from TREC Data. *Journal of Information Science*, 27(5), 303–310.

Keen, E.M. (1991). The Use of Term Position Devices in Ranked Output Experiments. *Journal of Documentation*, 47, 1–22.

Korfhage, R. (1997). *Information Storage and Retrieval.* New York: Wiley Computer Pub.

Lai, Y.S. and Wu, C.H. (2002). Meaningful Term Extraction and Discriminative Term Selection in Text Categorization via Unknown-Word Methodology. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1), 34–64.

Luhn, H.P. (1957). A Statistical Approach to the Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4), 309–317.

Luhn, H.P. (1958). The Automatic Creation of Literature Abstract. *IBM Journal of Research and Development*, 2(4), 159–165.

Meadow, C.T. (1992). *Text Information Retrieval System*. California: San Diego Academic Press.

Melucci, M. (1998). Passage Retrieval: A Probabilistic Technique. *Information Processing & Management*, 34(1), 43–68.

Ponte, J.M. and Croft, W.B. (1998). A Language Modeling Approach to Information Retrieval. In *Proceedings of 21st Annual International SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281). Melbourne, Australia: ACM.

Rasmussen, E. (1992). Clustering Algorithms. In W.B. Frakes and R. Baeza-Yates (Eds.), *Information Retrieval: Data Structures and Algorithms* (pp. 419–442). Englewood Cliffs, NJ: Prentice Hall.

Ro, J.S. (1988). An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II. On the Effectiveness of Ranking Algorithms on Full-Text Retrieval. *Journal of the American Society for Information Science*, 39(3), 147–160.

Robertson, A.M. and Willett, P. (1996). An Upperbound to the Performance of Ranked-Output Searching: Optimal Weighting of Query Terms Using a Genetic Algorithm. *Journal of Documentation*, 52, 405–420.

Robertson, S.E., Thompson, C.L., and Macaskill, M.J. (1986). Weighting, Ranking and Relevance Feedback in a Front-end System. *Journal of Information Science*, 12(2), 71–75.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., and Gatford, M. (1994). Okapi at TREC-2. In *Proceedings of The Second Text Retrieval Conference* (pp. 21–34). Gaithersburgh, MD: GPO.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. New York: Addison-Wesley.

Salton, G., Allan, J., and Singhal, A. (1996). *Information Processing and Management*, 32(2), 127–138.

Salton, G. and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5), 513–523.

Salton, G. and Yang, C.S. (1973). On the Specification of Term Values in Automatic Indexing. *Journal of Documentation*, 29(4), 351–372.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Information Retrieval. *Journal of Documentation*, 28, 11–21.

Sparck Jones, K. (1973). Indexing Term Weighting. *Information Storage and Retrieval*, 9, 619–633.

Umino, B. (1988). Some Principles of Weighting Methods Based on Word Frequencies for Automatic Indexing. *Library and Information Science*, 26, 67–88.

van Rijsbergen, C.J. (1977). A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*, 33(2), 106–119.

van Rijsbergen, C.J. (1979). *Information Retrieval*, 2nd ed. London: Butterworths.

Wilbur, W.J. (1993). Retrieval Testing with Hypergeometric Document Models: Global Term Weighting Approach. *Journal of the American Society for Information Science*, 44, 340–351.

Zobel, J. and Moffat, A. (1998). Exploring the Similarity Space. *ACM SIGIR Forum*, 32(1), 18–34.