To appear in the *International Journal of General Systems*
Vol. 00, No. 00, Month 20XX, 1–9

A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (tf-idf) Heuristic (and Variations Motivated by This Explanation)

Lukáš Havrlant[a,b] and Vladik Kreinovich[c*]
[a] *work performed at Department of Computer Science, Palacky University Olomouc, 17. listopadu 12, CZ-771 46 Olomouc, Czech Republic;*
[b] *currently at U Hřbitova 33, Opava, 74706, Czech Republic, Email: lukas@havrlant.cz;*
[c] *Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA*

(*Received 23 May 2014; accepted XX month 20XX*)

In document analysis, an important task is to automatically find keywords which best describe the subject of the document. One of the most widely used techniques for keyword detection is a technique based on the term frequency-inverse document frequency (tf-idf) heuristic. This techniques has some explanations, but these explanations are somewhat too complex to be fully convincing. In this paper, we provide a simple probabilistic explanation for the tf-idf heuristic. We also show that the ideas behind explanation can help us come up with more complex formulas which will hopefully lead to a more adequate detection of keywords.

Keywords: keywords; term frequency-inverse document frequency (tf-idf); probabilistic explanation

1. tf-idf: A Brief Reminder and Formulation of the Problem

How to find keywords: qualitative idea. Given a document, how can we identify its keywords? In many cases, this is easy: e.g., if we have a text which mentions Turing machines many times, then clearly the term "Turing machine" should be selected as a keyword. This does not mean, of course, that every word which occurs many times in a document is a meaningful keyword – e.g., words like "a" or "the" occurs many times in a document, but we should select them as keywords characterizing a given document – since they occur many times in every document.

So, to identify keywords, it is necessary to take into account not only how many times a given word t occurs in a given document d, but also how frequently the word t occurs in other documents.

How to find keywords: a (semi-empirical) algorithm. Information retrieval and text mining use a special term frequency-inverse document frequency (tf-idf) algorithm for automatic detection of keywords; see, e.g., Manning, Raghavan, and Schütze (2008); Rajaramann and Ullman (2011). The ideas behind this algorithm were first proposed in Jones (1972). This algorithm uses the following three numerical characteristics:

*Corresponding author. Email: vladik@utep.edu

- the number of time $\mathrm{tf}(t,d)$ that the word t occurs in a document d; this number is known as *term frequency*; and
- the total number N of documents in a given corpus D; and
- the number of documents $\mathrm{df}(t)$ which contain the given term t; this number is known as the *document frequency*.

Based on the last two characteristics, the algorithm computes the quantity $\mathrm{idf}(t,D) \stackrel{\mathrm{def}}{=} \ln \dfrac{N}{\mathrm{df}(t)}$ known as the *inverse document frequency*.

As keywords characterizing a given document d, we then select words t with the largest value of the product $\text{tf-idf}(t,d,D) \stackrel{\mathrm{def}}{=} \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$.

Remaining challenge. The tf-idf algorithm works reasonably well: in many cases, it leads to an adequate selection of keywords. What is not clear is why the above formula is so successful while other similar formulas (see, e.g., Manning, Raghavan, and Schütze (2008)) are not so successful.

There have been several attempts to provide a theoretical explanation for the success of tf-idf Heimstra (2000); Jones (1972); Manning, Raghavan, and Schütze (2008); Robertson (2004), but the resulting explanations are somewhat over-complicated and not very convincing.

What we do in this paper. In this paper, we provide a simple probabilistic explanation for the tf-idf heuristic.

This explanation motives some modifications of the original tf-idf formulas; we hope that these modifications will be useful too.

2. A Simple Probabilistic Explanation of tf-idf

Simplified model: main idea. Let us denote the total number of occurrences of the word t in the whole corpus D by $\mathrm{tf}(t,D)$. Let us consider a simplified model in which each of these occurrences is randomly assigned (with equal probability) to one of the N documents from the corpus D (and different occurrences are assigned independently from one another). In this model, the probability that each occurrence of the word t is assigned to a given document is equal to $\dfrac{1}{N}$.

Let us estimate the probability that this simplified model leads to the given number of occurrences. Let us estimate the probability p that after randomly (and independently) assigning all $n \stackrel{\mathrm{def}}{=} \mathrm{tf}(t,D)$ occurrences, the document d will contain $k \stackrel{\mathrm{def}}{=} \mathrm{tf}(t,d)$ occurrences.

The smaller this probability, the less probable it is that the text got k occurrences randomly, and thus, the more confident we are that the word t is important for the given document – i.e., that t is one of d's keywords.

Analysis of the simplified model and the resulting formula for the desired probability. Let us start with the case $k = 1$, when the document contains exactly one occurrence of the word t. To compute this probability, let us first estimate the probability that the assignment of the first of n words t placed this word into the given document d, and all the other $n-1$ assignments placed the corresponding word in other documents. The probability that, out of N documents, the first assignment is placed into the document d, is equal to $\dfrac{1}{N}$; the probability that each of the next $n-1$ assignments is placed in one of other $N-1$ documents is equal to

$\frac{N-1}{N} = 1 - \frac{1}{N}$. Since the assignments of different occurrences are independent, the resulting probability of this situation is

$$\frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{n-1}.$$

The overall probability that $k = 1$ comes from n such incompatible events: the event that the first occurrence landed up in d, the event that the second occurrence landed up in d, etc. Thus, the overall probability p that $k = 1$ is equal to the sum of n such terms, i.e., to

$$p = n \cdot \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)^{n-1}.$$

For $k = 2$, we can similarly compute the corresponding probability p: for each pair of occurrences, the probability that these two occurrences were placed in d and all $n - 2$ others were placed in other $N - 1$ documents is equal to

$$\left(\frac{1}{N}\right)^2 \cdot \left(1 - \frac{1}{N}\right)^{n-2}.$$

Thus, the probability p can be obtained by multiplying this probability by the total number $\binom{n}{2}$ of such pairs:

$$p = \binom{n}{2} \cdot \left(\frac{1}{N}\right)^2 \cdot \left(1 - \frac{1}{N}\right)^{n-2}.$$

Similarly, for a general k, for each k-tuple of occurrences, the probability that these k occurrences were placed in d and all $n-k$ others were placed in other $N-1$ documents is equal to

$$\left(\frac{1}{N}\right)^k \cdot \left(1 - \frac{1}{N}\right)^{n-k}.$$

Thus, the probability p can be obtained by multiplying this probability by the total number $\binom{n}{k}$ of such tuples:

$$p = \binom{n}{k} \cdot \left(\frac{1}{N}\right)^k \cdot \left(1 - \frac{1}{N}\right)^{n-k}.$$

Analysis of the problem and the resulting inequalities between k, n, and N. We are interested in the cases when the total number n of occurrences of the word t is much smaller than the total number N of documents in the corpus: $n \ll N$. Indeed, if n is approximately the same or greater than N – as is the case of such words as "a", "the", etc. – this means that the word t occurs in a large portion of documents and is, therefore, not typical for the given document d – so it cannot serve as one of its keywords.

We are also interested in the cases when the total number k of the occurrences of the term t in the given document is much smaller than its total number of occurrences n in the whole corpus of documents. Let us explain this requirement. Of course, by definition, k is always smaller than or equal to n. If k is of the same order as n, this means that on average, there are very few documents that contain this term. This can happen, for example, if an author introduced a new technical term in one paper and uses this term in another paper. However, in this case, it is not a good idea to use this new term as a keyword: one of the main purposes of the keyword is to make it to clear to the reader what this paper is about. From this viewpoint, using, as a keyword, a term which no one uses – and thus, most probably, no one understands – makes no sense. Thus, keywords are meaningful only if $k \ll n$.

Finally, for a word t to be a reasonable keyword for a document d, it should appear several times in the document: $1 \ll k$.

Summarizing, when we look for meaningful keywords, we should limit ourselves to cases when

$$1 \ll k \ll n \ll N.$$

The above inequalities help simplify the expression for the probability. Let us show how the above inequalities allow us to simplify the above expression for the probability p. Specifically, the above expression represents the probability p as the product of three factors: $\binom{n}{k}$, $\left(\frac{1}{N}\right)^k$, and $\left(1-\frac{1}{N}\right)^{n-k}$; we will show that the first and the third factors can simplified.

First, by using the expansion of the function $(1-x)^{n-k}$ in Taylor series, we get

$$\left(1-\frac{1}{N}\right)^{n-k} = 1-(n-k)\cdot\frac{1}{N}+\ldots$$

Since $n \ll N$, we have $n-k \ll N$, hence $(n-k)\cdot\frac{1}{N} \ll 1$ and so, in the first approximation,

$$\left(1-\frac{1}{N}\right)^{n-k} \approx 1.$$

To simplify an expression for $\binom{n}{k}$, let us use an explicit expression for the number of combinations:

$$\binom{n}{k} = \frac{n\cdot(n-1)\cdot(n-2)\cdots(n-k)}{1\cdot 2\cdots k}.$$

This can be equivalently described as

$$\binom{n}{k} = \frac{n^k}{k!}\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k}{n}\right).$$

Since $k \ll n$, we similarly get $1 - \dfrac{1}{n} \approx 1, \ldots, 1 - \dfrac{k}{n} \approx 1$ and therefore,

$$\binom{n}{k} \approx \frac{n^k}{k!}.$$

Since $k \gg 1$, we can use Stirling formula (see, e.g., Abramowitz and Stegun (2002)) for the factorial $k! \approx \left(\frac{k}{e}\right)^k$, so

$$\binom{n}{k} \approx \frac{n^k \cdot e^k}{k^k}.$$

Substituting these two approximate expressions for the factors into the formula for the probability p, we get an approximate formula

$$p \approx \frac{n^k \cdot e^k}{k^k} \cdot \left(\frac{1}{N}\right)^k.$$

From probabilities to their logarithms. The corresponding computations can be further simplified if we use logarithms of the probabilities instead of the probabilities themselves. Since logarithm is monotonic, the use of probabilities does not change which term is more probable and which is less probable; however, since $\ln(a \cdot b) = \ln(a) + \ln(b)$ and $\ln(a^k) = k \cdot \ln(a)$, the use of logarithms replaces multiplication with a computationally simpler addition operation, and raising to the power with a computationally simpler multiplication – this is why logarithms were invented in the first place.

Since the probability p is smaller than 1, its logarithm $\ln(p)$ is negative; to make it more convenient, let us consider its opposite $-\ln(p)$. From the above approximate formula, we conclude that

$$-\ln(p) \approx -k \cdot n + k \cdot \ln(k) - k + k \cdot \ln(N) = k \cdot \ln\left(\frac{N}{n}\right) + k \cdot (\ln(k) - 1).$$

Here, $k \gg 1$, so $\ln(k) \gg 1$. Thus, we arrive at the following formula:

Resulting formula.

$$-\ln(p) \approx k \cdot \ln\left(\frac{N}{n}\right) + k \cdot \ln(k).$$

The smaller the probability p, the larger this value and therefore, the more probable it is that the word t is one of the keywords describing the document d. Thus, as keywords describing a document, we should select all the terms t for which this expression is the largest.

Let us compare this formula with the tf-idf formula. The tf-idf formula has the form

$$k \cdot \ln\left(\frac{N}{\widetilde{n}}\right),$$

where $\widetilde{n}$ is the number of documents that contain the term t.

We are considering the case when the number n of occurrences of the word t is much smaller than the overall number of documents N. In this case, the probability that a document contains the word t can be estimated as $\frac{n}{N}$, and the number of such documents can be estimated as $N \cdot \frac{n}{N} = n$. The probability that a document contains two occurrences of the word t can be estimated as $\left(\frac{n}{N}\right)^2 \ll \frac{n}{N}$, and the number of such documents can be estimated as $N \cdot \left(\frac{n}{N}\right)^2 = n \cdot \frac{n}{N}$. Since $n \ll N$, this number is much smaller than the overall number of documents that contain the word t. So, most documents contain just one occurrence of the term t, and the overall number n of occurrences of the term t is approximately equal to the number $\widetilde{n}$ of the documents that contain t: $\widetilde{n} \approx n$. Hence,

$$-\ln(p) \approx k \cdot \ln\left(\frac{N}{\widetilde{n}}\right) + k \cdot \ln(k).$$

When $\frac{N}{n} \gg k$, we have $\ln(k) \ll \ln\left(\frac{N}{\widetilde{n}}\right)$, and therefore,

$$-\ln(p) \approx k \cdot \ln\left(\frac{N}{\widetilde{n}}\right).$$

This is exactly the tf-idf formula that we wanted to explain. Thus, we indeed get a simple probabilistic justification of the tf-idf formula.

Beyond explanation, towards a more accurate formula. The above analysis enables us not only to justify the *existing* semi-heuristic tf-idf formula, we can also provide a *new* formula which more accurately describes the probabilistic meaning and which, we hope, will be even more adequate in selecting keywords. Namely, instead of selecting keywords based on the tf-idf product expression, we should select keywords based on the value

$$\mathrm{tf}(t,d) \cdot \ln(\mathrm{nidf}(t,D)) + \mathrm{tf}(t,d) \cdot \ln(\mathrm{tf}(t,d)),$$

where the new measure of inverse document frequency is defined as

$$\mathrm{nidf}(t,D) \stackrel{\mathrm{def}}{=} \frac{N}{\mathrm{tf}(t,D)},$$

and, as before, $\mathrm{tf}(t,D)$ is the total number of occurrences of the term t in the whole corpus D of documents.

Comment. We can get an even more accurate description of the probability if we consider a more realistic (and, thus, more complex) probabilistic model.

3. A More Realistic Probabilistic Model and the Resulting Modification of tf-idf

Towards a more accurate model. In the above simplified model, we treated all the documents in the corpus equally. In practice, some documents are longer and some are shorter. Clearly, if a document is longer, it has a higher probability to

contain several occurrences of the given term t. Let us show how we can take this fact into account.

Resulting model. We want to take into account that different documents have different number of words. Let us denote the total number of words in a document d by $w(d)$; we will call this number the *length* of the document d.

Let W be the total number of words in all the documents in the given corpus D. Out of these W words, we have $\widetilde{n} = \mathrm{tf}(t, D)$ occurrences of each word t. Thus, the probability $p(t)$ that a randomly selected word is the occurrence of the word t is equal to the ratio $p(t) = \dfrac{\mathrm{tf}(t, D)}{W}$. The corresponding probabilistic model is straightforward: into each of W word locations, we place a term t with probability $p(t)$, and assignments corresponding to different locations are independent.

How probable it is that, as a result of this random assignment, in a document d with $w(d)$ words, we will get $\mathrm{tf}(t, d)$ words? The lower the probability of this result, the more probable it is that the word t is one of the keywords of the document d.

Analysis of the probabilistic model. After the above-described random assignment, the resulting number of occurrences of t can be computed as the sum $\mathrm{tf}(t, d) = x_1 + \ldots + x_{w(d)}$, where:

- $x_i = 1$ if the word at the i-th location is t and
- $x_i = 0$ if the word at the i-th location is different from t.

Since assignments corresponding to different locations are independent and identically distributed, the value $\mathrm{tf}(t, d)$ is thus equal to the sum of $w(d)$ independent identically distributed random variables. A document usually contains a reasonably large number of words; so, to describe the probability distribution of the value $\mathrm{tf}(t, d)$, we can use the Central Limit Theorem, according to which the probability distribution of the sum of many independent identically distributed random variables is close to Gaussian (normal); see, e.g., Sheskin (2011).

A normal distribution is uniquely determined by its mean μ and its variance $V = \sigma^2$. When we add independent random variables, their means add and their variances add. Thus, for the sum of $w(d)$ independent identically distributed random variables x_i, we get:

- $\mu = w(d) \cdot \mu_i$, where μ_i is the mean of each of the variables x_i, and
- $V = w(d) \cdot V_i$, where V_i is the mean of the variable x_i.

Each variable x_i has two possible values v_j:

- the value $v_1 = 1$ with probability $p_1 = p(t)$, and
- the value $v_0 = 0$ with the remaining probability $p_0 = 1 - p(t)$.

Thus,

$$\mu_i = \sum_j p_j \cdot v_j = p(t) \cdot 1 + (1 - p(t)) \cdot 0 = p(t).$$

Similarly,

$$V_i = \sum_j p_j \cdot (v_j - \mu_i)^2 = p(t) \cdot (1 - p(t))^2 + 0 \cdot (0 - p(t))^2 = p(t) \cdot (1 - p(t))^2 + (1 - p(t) \cdot p(t)^2.$$

The two terms in the right-hand side have a common factor, so

$$V_i = p(t) \cdot (1 - p(t)) \cdot ((1 - p(t)) + p(t)) = p(t) \cdot (1 - p(t)).$$

Thus,

$$\mu = w(d) \cdot \mu_i = w(d) \cdot p(t); \quad V = w(d) \cdot p(t) \cdot (1 - p(t)).$$

For normal distribution, possible values are values within the interval

$$[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma],$$

where k_0 is usually 2, 3, or 6. The larger $|k_0|$, the less probable it is for the corresponding value to appear. For a given value x, the corresponding value k_0 is determined by the equality $\mu \pm k_0 \cdot \sigma = x$, so $k_0 = \pm\dfrac{x - \mu}{\sigma}$, and $|k_0| = \dfrac{x - \mu}{\sigma}$.

For $x = \mathrm{tf}(t, d)$, the resulting ratio is equal to

$$\frac{\mathrm{tf}(t, d) - w(d) \cdot p(t)}{\sqrt{w(d) \cdot p(t) \cdot (1 - p(t))}}.$$

Let us simplify this formula. A keyword should occur much more frequently in this document than it occurs in the corpus in general. Thus, when we look for keywords, we are interested only in the words for which $\dfrac{\mathrm{tf}(t, d)}{w(d)} \gg p(t)$. For such words, $\mathrm{tf}(t, d) \gg w(d) \cdot p(t)$; thus, $\mathrm{tf}(t, d) - w(d) \cdot p(t) \approx \mathrm{tf}(t, d)$ and therefore, the above formula gets a simplified form

$$\frac{\mathrm{tf}(t, d)}{\sqrt{w(d) \cdot p(t) \cdot (1 - p(t))}}.$$

Also, as we have discussed earlier, as meaningful keywords, we cannot take words like "a" or "the" which occur frequently in all the documents. Thus, meaningful keywords should be relatively rare: we should have $p(t) \ll 1$. For such words, $1 - p(t) \approx 1$, and we get an even simpler formula for the resulting criterion:

$$\frac{\mathrm{tf}(t, d)}{\sqrt{w(d) \cdot p(t)}}.$$

Substituting the expression for $p(t)$ into this formula, we get the following final expression.

Resulting formula. As keywords corresponding to the document d, we should select words t for which the following value is the largest:

$$\mathrm{tf}(t, d) \cdot \sqrt{\frac{W}{\mathrm{tf}(t, D) \cdot w(d)}}.$$

Let us compare the new formula with tf-idf expression. The tf-idf formula corresponds to the case when we ignore the fact that different documents have different lengths, i.e., in effect, assume that all the documents have the same length.

If $w(d) = \text{const}$, then the ratio $\dfrac{W}{w(d)}$ is simply the total number N of the documents, and the above formula takes the form

$$\text{tf}(t,d) \cdot \sqrt{\frac{N}{\text{tf}(t,D)}} = \text{tf}(t,d) \cdot \sqrt{\text{nidf}(t,D)}.$$

This formula is very similar to tf-idf (to be more precise, it is similar to the modification of tf-idf that we described in the previous section); the main difference is that, instead of the logarithm of the inverse document frequency, we take the square root.

Acknowledgements

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

This work was done when L. Havrlant was visiting the University of Texas at El Paso, a visit supported by grant IGA 2014, no. PrF 2014 034, of Palacky University.

The authors are thankful to the anonymous referees for valuable suggestions.

References

Abramowitz, M., and I. Stegun. 2002. *Handbook of Mathematical Functions*, New York: Dover Publications.

Heimstra, D. 2000. "A probabilistic justification for using tf×idf term weighting in information retrieval", *International Journal of Digital Libraries* 3: 131–139.

Jones, K. S. 1972. "A statistical interpretation of term specificity and its application in retreival", *Journal of Documentation* 28 (1): 11–21; reprinted in 2004, 60 (5): 493–502.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*, New York: Cambridge University Press.

Rajaraman, A. and J. D. Ullman. 2011. *Mining of Massive Datasets*, Cambridge, UK: Cambridge University Press.

Roberston, S. 2004. "Understanding inverse document frequency: on theoretical arguments for idf", *Journal of Documentation* 60 (5): 503–520.

Sheskin, D. J. 2011. *Handbook of Parametric and Nonparametric Statistical Procedures*, Boca Raton, Florida: Chapman & Hall/CRC.