Reprinted from: *Journal of Documentation* **60** no. 5, pp 503–520

Understanding Inverse Document Frequency: On theoretical arguments for IDF

Stephen Robertson

Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB
UK
(and City University, London, UK)

Abstract

The term weighting function known as IDF was proposed in 1972, and has since been extremely widely used, usually as part of a TF*IDF function. It is often described as a heuristic, and many papers have been written (some based on Shannon's Information Theory) seeking to establish some theoretical basis for it. Some of these attempts are reviewed, and it is shown that the Information Theory approaches are problematic, but that there are good theoretical justifications of both IDF and TF*IDF in traditional probabilistic model of information retrieval.

1 Introduction

In 1972, Karen Spärck Jones published in the *Journal of Documentation* a paper called "A statistical interpretation of term specificity and its application in retrieval" (Sparck Jones, 1972). The measure of term specificity first proposed in that paper later became known as *inverse document frequency*, or IDF; it is based on counting the number of documents in the collection being searched which contain (or are indexed by) the term in question. The intuition was that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents, and the measure was an heuristic implementation of this intuition.

The intuition, and the measure associated with it, proved to be a giant leap in the field of information retrieval. Coupled with TF (the frequency of the term in the document itself, in this case, the more the better), it found its way into almost every term weighting scheme. The class of weighting schemes known generically as TF*IDF, which involve multiplying the IDF measure (possibly one of a number of variants) by a TF measure (again possibly one of a number of variants, not just the raw count) have proved extraordinarily robust and difficult to beat, even by much more carefully worked out models and theories. It has even made its way outside of text retrieval into methods for retrieval of other media, and into language processing techniques for other purposes[1].

One recurring theme in papers about IDF has been its heuristic nature. This has led many authors to look for theoretical explanations as to why IDF works so well. The number

of papers that start from the premise that IDF is a purely heuristic device and end with a claim to have provided a theoretical basis for it is quite startling (a few of these papers will be discussed below). The purpose of the present paper is to discuss the various approaches to this challenge and to demonstrate that we do actually have a good theoretical basis for IDF and explanation for its effectiveness.

2 The basic formula

Assume there are N documents in the collection, and that term t_i occurs in n_i of them. (What might constitute a 'term' is not of concern to us here, but we may assume that terms are words, or possibly phrases or word stems. 'Occurs in' is taken as shorthand for 'is an index term for', again ignoring all the difficulties or subtleties of either automatic indexing from natural language text, or human assignment of index terms.) Then the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

$$idf(t_i) = \log \frac{N}{n_i} \tag{1}$$

Actually this is not quite accurate – the original measure was an integer approximation to this formula, and the logarithm was specifically to the base 2. However, as will be seen below, the base of the logarithm is not in general important. Equation 1 is the most commonly cited form of IDF; some other forms will be indicated below.

3 Zipf's law

The original Sparck Jones paper made small theoretical claims for the measure, other than appealing to the intuition outlined above. However, some reference was made to the well-known Zipf law concerning term frequencies (Zipf, 1949). One version of Zipf's law is that if we rank words in order of decreasing frequency in a large body of text, and plot a graph of the log of frequency against the log of rank, we get a straight line – see Figure 1. Then it seems appropriate to put the terms into buckets by dividing the x-axis into equal intervals, which because of the straight line also correspond to equal intervals on the y-axis, as shown in the figure. This bucketting is essentially what the integer approximation used in the original IDF does.

This argument is not a very strong one. It is hard to see any formal justification for the above statement that 'it seems appropriate to . . . '. Besides, the term frequency used in Zipf's law (number of occurrences of the term in a body of continuous text) is not the same as the term frequency used in IDF (number of documents in which the term occurs, irrespective of how many times it occurs in each document). This distinction is an example of an *event space* problem which will be discussed further below.

4 Probabilities, logarithms and additivity

The fraction inside the logarithm in equation 1 looks like it might represent a probability (actually inverted). Thus we can consider the probability that a random document d would

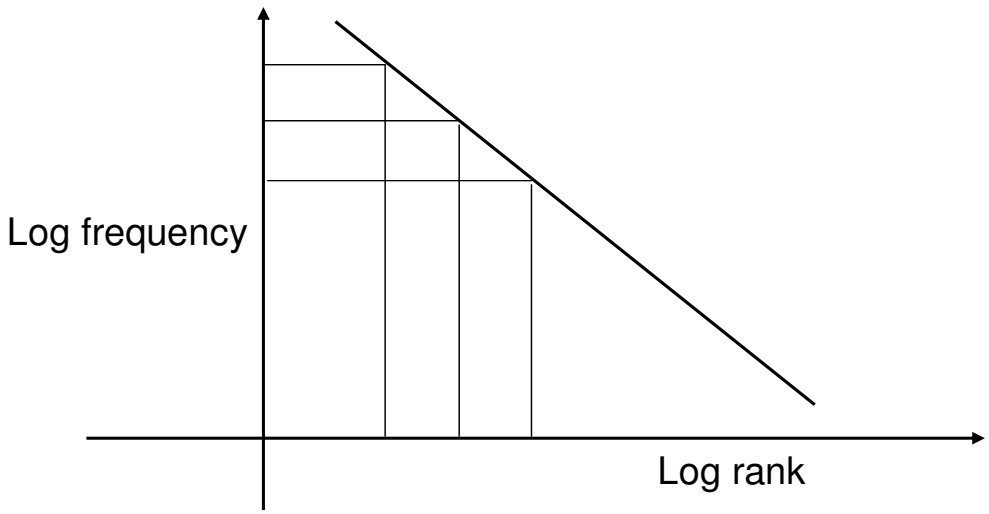

Figure 1: Zipf's law

contain the term (Robertson, 1972). This probability has an obvious estimate, namely the inverse of the fraction in the IDF equation:

$$P(t_i) = P(t_i \text{ occurs in } d) \approx \frac{n_i}{N}$$

In the light of this relation, we can reasonably redefine IDF in terms of the probability, and regard the observed IDF as an estimate of the true IDF:

$$idf(t_i) = -\log P(t_i) \tag{2}$$

(recall that $-\log x = \log \frac{1}{x}$).

One important feature of term weighting schemes is that it is frequently assumed that document scoring functions will be essentially additive. Thus if for example we have three query terms t_1, t_2, t_3, and we give them simple weights w_1, w_2, w_3, then a common simple scoring function would score a document by the sum of the weights of the query terms it contains. So a document containing all three terms would score $w_1 + w_2 + w_3$, while one containing only terms t_1 and t_3 would score $w_1 + w_3$. Clearly scoring functions can be considerably more complex than this, and we will encounter some below, but most are based on the addition of separate term components as a basic assumption.

We can combine the probabilistic interpretation of IDF and the idea of an addition-based scoring function very simply and elegantly: if we assume that the occurrences of different terms in documents are statistically independent, then addition is the correct thing to do with the logs. That is,

$$\begin{aligned} idf(t_1 \wedge t_2) &= -\log P(t_1 \wedge t_2) \\ &= -\log(P(t_1)P(t_2)) \\ &= -(\log P(t_1) + \log P(t_2)) \\ &= idf(t_1) + idf(t_2) \end{aligned}$$

Here $t_1 \wedge t_2$ represents the 'term' which is the Boolean *and* of t_1 and t_2. The same argument applies to any number of terms. Thus taking logs and then adding the weights is exactly the right thing to do. We may also note that if the reason for taking logs is to be able to add them, then the base of the logarithm does not matter.

Of course we do not really assume that the terms are statistically independent. Nevertheless, that simplifying assumption at least suggests strongly that we should use the log rather than (say) a linear function of the fraction N/n_i, which would be equally compatible with the original intuition.

5 Shannon's information theory

The $-\log p$ function of a probability suggests a connection with Shannon's theory of information and communication (1948). In the Shannon theory, the information capacity of a channel of communication is measured by the entropy formula:

$$H = -\sum_{i=1}^{n_m} p_i \log p_i \tag{3}$$

where p_i is the probability of receiving message m_i. It is assumed here that exactly one out of n_m possible messages m_i will be received, with known probabilities p_i. Note that because of this constraint, we know that

$$\sum_{i=1}^{n} p_i = 1$$

Note also that if the base of the logarithm is 2, the units of H are 'bits'. However, as indicated above, the base does not matter for most purposes.

The model can be extended to multiple messages coming over a channel, so that for example we could apply it to a succession of words, each of which is a message in its own right. The identification of this information capacity measure with entropy was made not by Shannon himself, but subsequently.

One possible interpretation (also not made by Shannon himself) of the measure is that each individual message carries a quantified amount of information

$$I_i = -\log p_i \tag{4}$$

Then the information capacity of the channel is defined as the *expected* amount of information from the next message. This definition gives exactly the equation 3 above:

$$H = \sum_i p_i I_i = -\sum_{i=1}^{n_m} p_i \log p_i$$

We might be tempted to link this interpretation to the IDF formula, by regarding IDF as measuring the amount of information carried by the term. Indeed, something along these lines has been done several times, including by the present author (Robertson, 1974). Some problems with this view of IDF are discussed below. First, it will be useful to define some other concepts associated with the Shannon entropy.

5.1 Messages and random variables

We can abstract the above Shannon definition of entropy by thinking about discrete random variables rather than messages. First, recall the definition of a random variable. We require an *event space* (a space of possible random events), together with a deterministic function of the outcome of each event. Then we require a probability measure of some kind. Ideally the probability measure is defined on the event space, and then the probability associated with any value of the random variable can be derived as the probability of the class of events which map to that value. (A class of events is just an aggregate event, whose probability is the sum of the probabilities of the elementary events in the class.) We may, however, define the probability measure only in terms of the values of the random variable.

In the Shannon case, the random event is just the arrival of a (the next) message. The outcome is a specific message m_i, with probability p_i. The event space is the space of all possible specific messages, and the random variable is just the identity of the message. (Note that if we want to deal with a succession of words rather than just a single word, we must redefine both the event space and the random variable; this is easy if we take a fixed window of n words, but hard if we want to consider messages of variable length.)

Given an event space and a discrete random variable X with values x_i, the entropy is defined exactly as in equation 3, where p_i is the probability $P(X = x_i)$ and n_m is now the number of distinct values of X. We can also define other quantities of interest: e.g. the mutual entropy of two random variables, the conditional entropy of one random variable given another, the mutual information of two variables, the Kullback-Leibler divergence between two probability distributions. All these are closely related (see e.g. Aizawa, 2003).

In general, however, these quantities are defined over entire probability distributions, and do not obviously relate to single (elementary or aggregate) events. There is in fairly common use a notion of 'pairwise [or pointwise] mutual information', which does relate to single events. But the transition from probability distribution measures like H to pointwise measures line I_i can be problematic – see the Appendix for an illustration.

5.2 IR interpretation

One problem with any application of such ideas to IR lies in the identification of the event space. We have a number of entities to consider in this environment – for example, documents, terms, queries. The discussion in Robertson (2004) (relating only to the first and the last of these) is an indication of the complexity of this issue. The essence of the problem is that it is hard to identify a single event space and probability measure within which all the relevant random variables can be defined. Without such a unifying event space, any process that involves comparing and/or mixing different measures (even a process as simple as adding IDFs for different terms) may well be invalid in terms of a Shannon-like formalism. In the obvious interpretation of equation 2, we should consider the event space as the space of documents.

We might instead think of a document as a list of terms present (ignoring duplication and TF at this point), and consider each term in turn as a single event/message. However, this view is problematic because the complete message (i.e. the whole document) is of variable length, and it is hard to see how to relate the term event-space to the document event-space, and recover a probability defined only in this latter space.

Alternatively, we might think of a document as a vector of V zeros and ones, representing

the presence or absence of each term in the vocabulary. Here we get over the variable length problem, because the length of the document vector is fixed as V, the size of the vocabulary. In effect we decompose the single event which is a document into the cross product of V other spaces, each with only two elements. Cross-product event spaces are however problematic, for reasons discussed in Robertson (2004). But without going into those technicalities, we can identify one new problem: we have introduced a new set of probabilities: for element i there is not only the probability of a one (p_i), but also the probability of a zero ($1-p_i$). Thus somewhere in our weights we should have quantities like $-\log(1-P(t_i))$, representing the absence of a term. This might not matter, since typically $P(t_i)$ is small, so $(1-P(t_i))$ is close to 1 and the log is close to zero.

But there is a more serious problem. When we search using weighted terms, we typically take the query terms and assign weights to them, but ignore all the other terms in the vocabulary. It is hard to see how such a practice could make sense in a Shannon-like model: every term in the document must be assumed to carry information as well as any other. That is, the presence of a term t_i would carry $-\log P(t_i)$ amount of information irrespective of whether or not it is in the query. There is nothing to link the amount of information to the specific query. So we would have no justification for leaving it out of the calculation.

Nevertheless, the similarity of the usual IDF formulation to a component of entropy has stimulated other researchers to try to make connections, sometimes somewhat differently from the link suggested above. Two of these efforts are discussed briefly here.

Aizawa

In this paper (Aizawa, 2003), the author juggles with three different event spaces: the space of documents, the space of query terms and the space of terms in document texts. This leads him to make several 'restrictive assumptions' about the relationships between the relevant probability distributions, and to claim that these assumptions "represent the heuristic that tf-idf employs". He also encounters the problem of single events mentioned above – he uses a 'conditional entropy' which is conditional on an event rather than on a random variable. This in effect introduces yet more event spaces, for example the space of documents which contain a particular term. Probabilities in one event space are estimated using data from another event space. While the analysis may shed light on some of the complexities involved, it is difficult to see it as an explanation of the value of IDF. There are certainly other ways to interpret 'the heuristic that tf-idf employs', as discussed below.

Siegler and Witbrock

This much shorter paper (Siegler and Witbrock, 1999) defines both queries and documents as mappings of words onto probabilities – in other words, the event space is the space of words but with two different probability measures (to which they confusingly give the same notation). Again, both conditional entropy and mutual information are defined with respect to a conditioning event (=word) rather than to a conditioning random variable. They appear to derive the result that the IDF value for a term is exactly the mutual information of the document collection and the term; however, the derivation contains some obvious errors of algebra.

Thus despite the $-\log P$ form of the traditional IDF measure, any strong relationship between it and the ideas of Shannon's information theory is elusive. It seems unlikely that

such a relationship will give us a strong explanation, validation or justification for the use and value of IDF (or TF*IDF) as a term-weighting method in information retrieval.

Another interpretation: Papineni

Papineni (2001) makes some appeal to information theory ideas (maximum entropy, Kullback-Leibler distance, mutual information), but his main focus is on a proof of optimal performance. The question he asks is: given a document as query, what is the best classifier to retrieve that same document. The answer (under an independence assumption) is to give the terms IDF weights.

Although Papineni makes no reference to the fact (and indeed starts with claims about the heuristic nature of IDF), his model is equivalent to a much-simplified version of the older relevance-based probabilistic model described in the next section. Essentially, it assumes that (a) for any query there is only one relevant document, and (b) all query terms are guaranteed to be in the document. Neither of these assumptions is necessary. Where Papineni's argument does seem to have some benefits (but maybe these do depend on the simplification) is in (a) giving some handle on dealing with dependencies between terms (via bigrams), and (b) distinguishing between the term weight and the gain from including the term as a feature in the classifier. (However, this distinction was also made earlier in the IR field – see Robertson, 1990.)

6 The classical probabilistic model for IR

The interpretation of IDF as a function of a probability suggests associating it with (one of) the probabilistic approaches to information retrieval. There is a range of different approaches to IR that can be described broadly as 'probabilistic', including a number of comparatively recent developments using machine learning and/or statistical language modelling techniques. However, we will start this discussion with the probability ranking principle and search term weighting based on an explicit relevance variable. This general approach is discussed by van Rijsbergen (1979); the series of models described below (including those referred to here as RSJ and BM25) is developed in Sparck Jones, Walker and Robertson (2000). It is not yet quite obvious how IDF might relate to a relevance variable, but this will emerge below.

The probability ranking principle (Robertson, 1977) states that for optimal performance, documents should be ranked in order of probability of relevance to the request or information need, as calculated from whatever evidence is available to the system. Essentially it makes an explicit connection between the ranking of documents which the system may give the user, and the effectiveness of the system in terms of the conventional experimental measures of performance like recall and precision.

6.1 Relevance weights

The relevance weighting model (Robertson and Sparck Jones, 1976), referred to as RSJ, shows that under some simple assumptions, this probability of relevance ordering can be achieved by assigning weights to query terms, and scoring each document by adding the weights of the query terms it contains. (In this version of the model, the only evidence about the documents that is used is the presence or absence of terms; term frequency in the document

will be discussed below.) The term weight may be defined as follows. First we define two probabilities for term t_i:

$$\begin{aligned} p_i &= P(\text{document contains } t_i | \text{document is relevant}) \\ q_i &= P(\text{document contains } t_i | \text{document is not relevant}) \end{aligned}$$

Then the term weight is:

$$\text{RSJ weight} \quad w_i^{(1)} = \log \frac{p_i(1-q_i)}{(1-p_i)q_i} \tag{5}$$

Next, we consider how to estimate the two probabilities and therefore the term weight. We initially assume for this purpose that we know which documents are relevant and which are not. This assumption is of course not only unrealistic, but also makes a nonsence of the search process. Below we consider the cases where we have either partial relevance information (we know of some relevant documents), or none at all. If we have, as before, a total of N documents of which n_i contain the terms, and further R out of the N are relevant and r_i relevant documents contain the term, then the obvious estimates are:

$$p_i \approx \frac{r_i}{R}, \qquad q_i \approx \frac{n_i - r_i}{N - R} \tag{6}$$

However, because of the form of Equation 5 (essentially log-odds rather than straight probabilities), a modification to these estimates is appropriate. The usual modification may be justified in various ways, not discussed here (there is one such in (Robertson and Sparck Jones, 1976)), but the effect is to use the following estimate of the RSJ weight, where the 0.5 added to each of the components can be seen as a smoothing correction:

$$w_i = \log \frac{(r_i + 0.5)(N - R - n_i + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)} \tag{7}$$

The assumptions of this relevance weighting model include assumptions of statistical independence between terms, somewhat like but not quite the same as the independence assumption given in section 4 above. The difference is that here terms are assumed to be independent within the set of relevant documents, and again within the set of non-relevant documents. In general, they will not be independent in the collection of documents taken as a whole – in fact good query terms will in general be positively correlated in the collection as a whole, even if independent given relevance[2]. This positive correlation makes these assumptions just slightly more realistic than the one in section 4, though not very much so. They serve to characterise the relevance weighting model as a 'naïve Bayesian' model in the current terminology of the machine learning literature.

Given these independence assumptions, the RSJ weights are additive. That is, the best use we can make of the information about term presence and absence information, according to the probability ranking principle, is to weight the terms as above, score the documents as the sum of weights of the terms present in the documents, and then rank the documents in score order. This is the strong result of the relevance weighting theory – in fact it is a form of proof of optimal performance, very much more general than that of Papineni discussed above.

6.2 Query terms and others

In the relevance weighting theory, we normally apply the above weights to query terms only. This corresponds to the assumption that (in the absence of evidence to the contrary) we

may take the value of the RSJ weight of equation 5 to be zero for non-query terms. That is, non-query terms are assumed not to be correlated with relevance, and this implies giving them zero weight, so that leaving them out of the calculation is entirely correct. This is not necessarily a realistic assumption, but it does make clear one large difference between the probabilistic model and a Shannon-like view. That is, in the probabilistic model, a simply-expressed assumption clearly ties in with the practice of restricting a search to query terms.

In this paper, we concentrate on the term weighting issue. However, it is also worth mentioning that with partial relevance information (as discussed in the next section) we can avoid the assumption about non-query terms and instead consider query expansion. That is, we can look for evidence of terms that are correlated with relevance, and to use them to expand the original query. Thus the relevance weighting theory provides a more-or-less principled reason either to restrict the matching to query terms, or to choose some additional terms for the query. (As noted above, the choice of terms to add to the query should not be determined by the optimal term weight for searching, but requires a related but different measure of the effect of adding such a term to the query (Robertson, 1990), what Papineni refers to as 'gain' (Papineni, 2001).

6.3 Little relevance information

Suppose now that we have little relevance information – the user either knows about in advance, or has seen and judged, just a few documents, and has marked R as relevant. Then the obvious way to estimate p_i is as in equation 6 or the smoothed form embedded in equation 7 (r_i is now the number of documents *known* to be relevant that contain the term). The appropriate estimate of q_i is somewhat less obvious – we could base it on the number of documents *known* to be non-relevant, but this is likely to be a very small number. On the other hand, we usually know or can be sure that almost all of the documents in the collection are non-relevant (that is the usual information retrieval situation). So if we take the complement of the known relevant documents, i.e. $(N - R)$ documents, to be non-relevant, this will give a somewhat biased but probably a much more reliable estimate. This complement method has been shown to work well. Then the estimation formula for w_i can remain exactly as in equation 7, with R and r_i representing known relevant documents only. The 0.5 correction is particularly appropriate for small samples.

6.4 No relevance information

What could we do if we have no relevance information? Concerning the probability p_i, relating to relevant documents, at first view we have no information at all. On that basis, probably the best we can do is to assume a fixed value p_0 for all the p_i (Croft and Harper, 1979). Concerning q_i, however, we can make the same assumption made in the previous paragraph – that the collection consists to a very large extent of non-relevant documents. An extreme version of the complement method described is that for the purpose of estimating q_i, we assume that *all* documents are non-relevant. Then q_i is estimated from the entire collection.

These assumptions, applied to equation 5 with the simple estimates of equation 6, yield the following:

$$w_i = C + \log \frac{N - n_i}{n_i} \tag{8}$$

where $C = \log \frac{p_0}{1-p_0}$. If instead of the estimates of equation 6 we use equation 7, setting $R = r = 0$, we get

$$w_i = \log \frac{N - n_i + 0.5}{n_i + 0.5} \tag{9}$$

Here the result of setting $R = r = 0$ with the point-5 correction is to set $p_0 = 0.5$ and therefore $C = 0$.

These two versions now look very much like IDF measures. The $-n_i$ on the top of the fraction makes them a little different from the original equation 1, but this difference is of little practical importance in most situations, because n_i is typically much smaller than N. Similarly the 0.5 smoothing correction is generally of small effect.

However, the $-n_i$ on top does cause some strange effects. In the (rare) case of a term which occurs in more than half the documents in the collection, either of the above two formulae defines a negative weight. This is a somewhat odd prediction for a query term – that its presence should count against retrieval. The original formula would simply give a small positive weight. This problem is investigated by Robertson and Walker (1997), who argue that the fixed p_i is the source of the problem, and is not reasonable for frequently occurring terms. They show that a modification of this assumption, in which p_i also increases with increasing frequency in the collection, avoids this anomalous behaviour by removing the n_i from the numerator in formulae 8 and 9. Thus equation 8 is replaced by

$$w_i = C + \log \frac{N}{n_i}$$

and equation 9 by

$$w_i = \log \frac{N + 0.5}{n_i + 0.5}$$

6.5 IDF as an RSJ weight

It is now apparent that we can regard IDF as a simple version of the RSJ weight, applicable when we have no relevance information. This is in fact a strong justification for IDF. It appeals to no notions of information theory, but only to a simple naïve Bayes model of retrieval together with some further plausible simplifying assumptions. The naïve Bayes model makes explicit use of a relevance variable; although the further simplifying assumptions render this variable implicit in IDF, we can still regard IDF as a direct measure of probability of relevance. This characteristic would seem to make the relevance-weight justification of IDF considerably stronger than any of the information-theory justifications considered.

7 TF and language models

Neither the original Sparck Jones proposal for IDF, nor the relevance weighting model of Robertson and Sparck Jones, make any use of within-document term frequency information (TF). So far, we have made no arguments for the use of TF, nor of how TF might be combined with IDF, as in the TF*IDF-style measures mentioned at the beginning of this paper. TF*IDF-style measures emerged from extensive empirical studies of combinations of weighting factors, particularly by the Cornell group (Salton and Yang, 1973).[3]

To understand or explain the effectiveness of these measures, we need some justification not just for the TF factor itself, but also for why one should want to multiply the log-like IDF weight by the TF-based component.

In what follows, we denote by tf_i the frequency (number of occurrences) of term t_i in the document under consideration. If we need to talk about multiple documents, we will refer to $tf_{i,j}$ for t_i in document d_j. Also dl or dl_j will refer to the length of the document – that is the number of term-positions, so that (summing over all the terms in the document or the vocabulary)[4]:

$$dl = \sum_i tf_i$$

7.1 Logs and additivity revisited

The very simple independence argument given in section 4 can be applied also to multiple occurrences of the same term (as well as to occurrences of different terms). If we assume that successive occurrences of a single term are independent of each other (that is, the number of times that we have seen this term already in this document does not affect the probability that we will see it again), then we can simply add another w_i for each time we see term t_i. Then when we cumulate over the complete document, we would get tf_i occurrences of each term t_i, so we should simply multiply the IDF weight w_i by the number of occurrences tf_i.

This simple argument conceals a number of problems. The major one is that we are no longer operating in the same event space: instead of documents, we are looking at term-positions in the text of each document. The event of seeing a particular term now relates to a particular term position instead of the document. Two consequences of this problem are (a) that we would now need to take account somehow of the document length, and (b) that the original basis on which we regarded $\frac{n_i}{N}$ as a probability has now gone. This latter issue seems to be fatal – it seems like we have pulled the rug out from under the whole idea of IDF.

One possible way to get away from this problem would be to make a fairly radical replacement for IDF (that is, radical in principle, although it may be not so radical in terms of its practical effects). We could replace n_i by $\sum_j tf_{i,j}$ and N by $\sum_j dl_j$, thus transferring the probability from the event space of documents to the event space of term positions in the concatenated text of all the documents in the collection. Then we have a new measure, called here *inverse total term frequency*:

$$ittf(t_i) = \log \frac{\sum_j dl_j}{\sum_j tf_{i,j}} \tag{10}$$

This could again be interpreted as the log of an inverse probability, but a probability in the term-position event space rather in than the document event space. It is sometimes referred to as inverse collection frequency, but as IDF has also been referred to as a collection frequency weight, in particular in the original Sparck Jones paper, this terminology is avoided here.

On the whole, experiments with inverse total term frequency weights have tended to show that they are not as effective as IDF weights. Church and Gale (1995) show how the two measures differ significantly on exactly the terms in which we might be interested.

It will be seen below that we can preserve the document event-space in which IDF seems to fit naturally, but nevertheless introduce a TF component.

7.2 Other formulations of TF*IDF

We have already seen that one of Aizawa's (2003) concerns was with TF, as an example of an attempt to use information theory for a derivation of a TF*IDF weighting scheme – and we have already seen some problems with such a formulation.

Joachims

The approach taken here (Joachims, 1997) appeals not to information theory, but to naïve Bayes machine learning models, as used in the relevance weighting theory (Joachims' task is not the usual ranked retrieval task, but a categorisation task). He confronts the event space problem in part by redefining IDF in a somewhat similar fashion to the inverse total term frequency measure discussed above (it is not quite the same, because he does not concatenate the texts of all the documents: instead he takes a measure based on term occurrences in each document and combines this measure across documents in a different way). He also replaces the usual log function of IDF with a square-root function. Thus the version of 'IDF' in his final formulation is quite significantly different from traditional IDF. On the way he has to make various assumptions, like that all the classes in his categorisation task have approximately the same number of documents (this is reminiscent of Papineni's assumption that there is just one relevant document per query).

Roelleke

Roelleke (2003) makes an analysis of IDF which starts off, somewhat confusingly, with a proposal to consider the quantity $\frac{idf}{maxidf}$ as if it were a probability (it is difficult to see any kind of event space in which this would look like either a probability or an estimate of one). He does, however, go on to consider two separate cross-product event spaces, based on similar considerations to those discussed above: $D \times T$ (documents from the collection and terms from the vocabulary, where the event is "term occurs in document") and, for each document, $L \times T$ (locations, i.e. term positions, in the document and terms from the vocabulary, where the event is "this term occurs in this position"). Clearly TF has some meaning in this latter space but not in the former. Roelleke draws some parallels between these two spaces, and devises an alternative to IDF which is based on the latter space; however, the alternative again looks rather different from IDF. He does not present an analogue to a TF*IDF formula.

Language models

There has been a great deal of work recently in the application of statistical language models (LM) to information retrieval (see e.g. Croft and Lafferty, 2003). In a statistical language model, each successive word position (in a document or query) is regarded as a slot to be filled with a word; a generative process is envisaged, in which these slots are filled on the basis of probabilities over the vocabulary. Thus these models operate in the term-position event space identified above, rather than the document event space (though there may be a separate language model for each document). In the original applications of these models (such as speech recognition, where they have been very successful), the probability distribution for a slot would typically depend on some small amount of history, normally what the previous few words were. However, much of the LM work in IR has been based on 'unigram' models, where the probabilities for each position in a document are assumed to be independent. (It is important to realise nevertheless that the LM approach still thinks of the sequence of term slots or positions, even if the probabilities are assumed to be independent.)

The relationship between such models and the matters discussed in this paper is a little unclear. Standard simple language models in IR (e.g. Ponte and Croft, 1998) do not make explicit use of either IDF or TF, but nevertheless achieve a similar effect to a TF*IDF measure. It seems possible that they are closer in spirit to the inverse total term frequency measure

defined above, since they are operating in a similar event space. Another difference is that the simple language models make no direct reference to relevance. However, some more recent work in this field does have an explicit relevance variable (e.g. Lavrenko and Croft, 2001).

The 2-Poisson model and the BM25 weighting function were formulated before the language modelling idea came along. However, they may be recast as an elementary form of language model. This view is explored in the next section.

8 TF in the relevance weighting model

Given the relative success, reported above, of an explanation of IDF based on the original relevance weighting model, it makes sense to look for an explanation of TF*IDF based on an extension of the relevance weighting model that accommodates TF. The weighting function known as Okapi BM25 (it was developed as part of the Okapi experimental system (Robertson, 1997), and was one of a series of Best Match functions) is probably the best-known and most widely used such extension.

8.1 Eliteness and the 2-Poisson model

The following is a brief account of the derivation of BM25. It does not attempt to give full technical detail. The original argument was presented by Robertson and Walker (1994), and an analysis of the various components is given by Sparck Jones et al. (2000). The present discussion makes some use of the more recent idea of language modelling as discussed above.

First, we assume that all documents are the same length – this assumption will be dropped below. Next we assume (this is the central assumption of the model) that each term (word) represents a concept; and that a given document is either 'about' the concept or not. This property is described as *eliteness*: the term is elite in the document or not. This terminology, as well as the statistical model described below, is taken from Bookstein and Swanson (1974) and Harter (1975). Eliteness is a hidden variable – we cannot observe it directly. However, we then assume that the text of the document is generated by a simple unigram language model, where the probability of any term being in any given position depend on the eliteness of this term in that document.

If we take all the documents for which this particular term is elite, then we can infer the distribution of within-document frequencies we should observe. If all documents are the same length, then the distribution will be approximately Poisson. If we take instead all the documents for which this term is *not* elite, we will again see a Poisson distribution (presumably with a smaller mean). But of course we cannot know eliteness in advance; so if we consider the collection as a whole, we should observe a mixture of two Poisson distributions. This two-Poisson mixture is the basic Bookstein/Swanson/Harter model for within-document term frequencies.

Eliteness thus provides the bridge between the event-space of documents (or rather the cross-product of documents in the collection and terms in the vocabulary) and the event-space of term positions, which gives us term frequencies. For each term, eliteness is a property of the document, but determines the properties of the term positions in the document.

Finally, we have to make the connection with relevance. Relevance is a property at the document level, so the connection is with eliteness rather than with term occurrences; but now we have to make the bridge between query and document terms. But we know how to do that, having done it in the original relevance weighting model. So now we re-use the relevance

weighting model, only applying it to query-term eliteness rather than to query-term presence or absence.

From the above arguments, we can formulate a weighting scheme which involves the following five parameters for each query term:

- the mean of the Poisson distribution of within-document term frequencies for elite documents;
- ditto for non-elite documents;
- the mixing proportion of elite and non-elite documents in the collection;
- the probability of eliteness given relevance; and
- the probability of eliteness given non-relevance.

The full equation is given in Robertson and Walker (1994). The problem with such a scheme is that all these parameters (five per query term) would need to be estimated, and both eliteness and (in general) relevance are hidden variables. This makes the full equation almost unusable. However, it is possible to simplify it to the point where it becomes feasible to think of using it. At the same time, we introduce a normalisation to deal with the fact that documents vary in length.

8.2 BM25

The BM25 weighting function is based on an analysis of the behaviour of the full eliteness model under different values of the parameters. Essentially, each term would have a full eliteness weight, that is, a weight that a document would gain if we knew that the term was elite in that document. If we do not know that, but have a TF value which provides probabilistic evidence of eliteness, we should give partial credit to the document. This credit rises monotonically from zero if $tf = 0$ and approaches the full eliteness weight asymptotically as tf increases. In general the first occurrence of the term gives most evidence; successive occurrences give successively smaller increases.

BM25 first estimates the full eliteness weight from the usual presence-only RSJ weight for the term, then approximates the TF behaviour with a single global parameter k_1 controlling the rate of approach. Finally it makes a correction for document length.

The document-length normalisation is not really relevant to the present discussion; however, it is presented for completeness. On the assumption that one reason for a document to be long is verbosity on the part of the author (which would suggest simple document-length normalisation of TF), but that a second reason is the topic coverage of the document (which would suggest no normalisation), BM25 does partial normalisation. The degree of normalisation is controlled by a second global parameter b. In order to make the normalisation relatively independent of the units in which document length is measured, it is defined in terms of the average length of documents in the collection.

The resulting formula can be expressed thus:

$$\text{BM25 weight} \quad w_i = f(tf_i) * w_i^{(1)} \tag{11}$$

where

$w_i^{(1)}$ is the usual RSJ weight,
$f(tf_i) = \frac{(k_1+1)tf_i}{K+tf_i}$
$K = k_1((1-b) + b * \frac{dl}{avdl})$
dl and $avdl$ are the document length and average document length respectively
k_1 and b are global parameters which are in general unknown, but may be tuned on the basis of evaluation data.

8.3 TF*IDF as a BM25 weight

It is clear that the formula given in equation 11 expresses BM25 as a TF*IDF weighting scheme. The first component is a TF component (although it has the important characteristic of being highly non-linear, and as a consequence, defining a ceiling to the contribution of any one term to the overall document score). The second component is the RSJ weight, which as we have seen may make use of relevance information if it exists, but reduces to an IDF measure in its absence.

Thus the theoretical argument behind BM25 can be seen to justify TF*IDF, or in some sense to explain why it works as well as it does, in just the same way that the relevance weighting theory explains IDF on its own.

9 Conclusions

The basic search term weighting formula known as IDF, proposed by Sparck Jones on heuristic grounds in 1972, has proved extraordinarily robust. It remains at the core of many, if not most, ranking methods used in search engines. In some more recent developments (specifically the language models) it is no longer an explicit component – in these models, a similar effect is achieved by somewhat different means. So IDF may in future disappear as an explicit component, but nevertheless live on in spirit.

IDF has also proved a magnet for researchers who feel inspired to replace what they perceive as an heuristic with some reasonably-constituted theoretical argument, in order to 'explain' why it is that IDF works so well. Such contributions often start with the assertion that IDF is *just* an heuristic, with no known or understood theoretical basis.

Such a theoretical explanation or justification for IDF is not completely straightforward. Although it is superficially easy, as a first step, to interpret IDF as a probabilistic function, one major problem area lies in the definition of the appropriate event space for the probabilities concerned. This problem undermines many attempted explanations. One fertile source of attempts has been Shannon's theory of information, in which IDF is usually interpreted more specifically, as a particular kind of probabilistic function: a measure of the amount of information carried by a term. But again, such attempts run into serious problems.

However, there is a relatively simple explanation and justification of IDF in the relevance weighting theory of 1976. This extends to a justification of TF*IDF in the Okapi BM25 model of 1994. IDF is simply neither a pure heuristic, nor the theoretical mystery many have made it out to be. We have a pretty good idea why it works as well as it does.

Acknowledgements

I am very grateful to Karen Spärck Jones for many useful comments and discussions.

Notes

[1] The history and range of influence of IDF have been described in other places (e.g. Harman, forthcoming).

[2] The association of each term separately with relevance induces an association between terms in the whole collection, i.e. when relevance is ignored.

[3] All of the early work was conducted on short document surrogates (e.g. short catalogue records or scientific abstracts) rather than full-text documents. In these conditions, TF is not so important as it subsequently became with full-text documents. Furthermore the document length component of a TF measure was generally unimportant, since typically document surrogates in a given collection varied little in length. The critical importance of good document length normalisation emerged only in the 1990s.

[4] The actual definition of document length can be somewhat more complicated, depending on such factors as stopwords or the indexing of phrases, and can be measured in a variety of ways, but these issues will be ignored in the present discussion.

References

Aizawa, A. (2003), "An information-theoretic perspective of tf-idf measures", *Information Processing and Management*, Vol. 39, pp. 45–65.

Bookstein, A. and Swanson, D. R. (1974), "Probabilistic models for automatic indexing", *Journal of the American Society for Information Science*, Vol. 25, pp. 312–319.

Church, K. and Gale, W. (1995), "Inverse Document Frequency (IDF): a measure of deviations from Poisson", in D. Yarowsky and K. Church (Eds), *Third Workshop on very large corpora*, ACL, MIT, pp. 121–130.

Croft, W. B. and Lafferty, J. (Eds) (2003), *Language Modelling for Information Retrieval*, Kluwer.

Croft, W. and Harper, D. (1979), "Using probabilistic models of information retrieval without relevance information", *Journal of Documentation*, Vol. 35, pp. 285–295.

Gray, R. M. (1990), *Entropy and Information Theory*, Springer Verlag.

Harman, D. (forthcoming), "The history of idf and its influences on IR and other fields", in J. Tait (Ed.), *??*, ??, p. ??

Harter, S. P. (1975), "A probabilistic approach to automatic keyword indexing (parts 1 and 2)", *Journal of the American Society for Information Science*, Vol. 26, pp. 197–206 and 280–289.

Joachims, T. (1997), "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", in D. H. Fisher (Ed.), *Proceedings of ICML-97, 14th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, US, Nashville, US, pp. 143–151.

Lavrenko, V. and Croft, W. B. (2001), "Relevance-based language models", in W. B. Croft, D. J. Harper, D. H. Kraft and J. Zobel (Eds), *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 120–128.

Papineni, K. (2001), "Why inverse document frequency?", *Proceedings of the North American Association for Computational Linguistics*, NAACL, pp. 25–32.

Ponte, J. M. and Croft, W. B. (1998), "A language modeling approach to information retrieval", in W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson and J. Zobel (Eds), *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 275–281.

Robertson, S. (2004), "On event spaces and probabilistic models in information retrieval", *Information Retrieval*, Vol. 7?, pp. to appear. Presented at SIGIR 2002 Workshop on Mathematical/Formal Methods in Information Retrieval; submitted to *Information Retrieval*.

Robertson, S. E. (1972), "Term specificity [letter to the editor]", *Journal of Documentation*, Vol. 28, pp. 164–165.

Robertson, S. E. (1974), "Specificity and weighted retrieval [documentation note]", *Journal of Documentation*, Vol. 30 No. 1, pp. 41–46.

Robertson, S. E. (1977), "The probability ranking principle in information retrieval", *Journal of Documentation*, Vol. 33, pp. 294–304.

Robertson, S. E. (1990), "On term selection for query expansion", *Journal of Documentation*, Vol. 46, pp. 359–364.

Robertson, S. E. (1997), "Overview of the Okapi projects [introduction to special issue of Journal of Documentation", *Journal of Documentation*, Vol. 53, pp. 3–7.

Robertson, S. E. and Sparck Jones, K. (1976), "Relevance weighting of search terms", *Journal of the American Society for Information Science*, Vol. 27, pp. 129–146.

Robertson, S. E. and Walker, S. (1994), "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval", in W. B. Croft and C. J. van Rijsbergen (Eds), *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag, pp. 345–354.

Robertson, S. E. and Walker, S. (1997), "On relevance weights with little relevance information", in N. J. Belkin, A. D. Narasimhalu and P. Willett (Eds), *SIGIR'97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 16–24.

Roelleke, T. (2003), "A frequency-based and a poisson-based definition of the probability of being informative", in J. Callan, G. Cormack, C. Clarke, D. Hawking and A. Smeaton

(Eds), *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, pp. 227–234.

Salton, G. and Yang, C. S. (1973), "On the specification of term values in automatic indexing", *Journal of Documentation*, Vol. 29, pp. 351–372.

Shannon, C. E. (1948), "A mathematical theory of communication", *The Bell System Technical Journal*, Vol. 27, pp. 379–423 and 623–656.

Siegler, M. and Witbrock, M. (1999), "Improving the suitability of imperfect transcriptions for information retrieval from spoken documents", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE Press, Piscataway, NJ, pp. 505–508.

Sparck Jones, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, Vol. 28, pp. 11–21.

Sparck Jones, K., Walker, S. and Robertson, S. E. (2000), "A probabilistic model of information retrieval: development and comparative experiments", *Information Processing and Management*, Vol. 36, pp. 779–808 (Part 1) and 809–840 (Part 2).

van Rijsbergen, C. J. (1979), *Information retrieval*, Butterworths, London, U.K. (Second edition).

Zipf, G. K. (1949), *Human behavior and the principle of least effort*, Addison-Wesley, Reading, Ma.

Appendix: Entropy and pointwise information

In section 5, we defined the Shannon notion of information capacity of a channel (entropy), and the apparently associated pointwise information measure for each separate message m_i. Equations 3 and 4 are reproduced here for convenience:

$$H = -\sum_{i=1}^{n_m} p_i \log p_i \tag{3}$$

$$I_i = -\log p_i \tag{4}$$

As noted, the definition of H can be interpreted as the expected value of I_i. That is, mathematically speaking H *is* the expected value of I_i; but in order to make the interpretation, we have to accept that it makes sense to measure the amount of information in an individual message, and equate the capacity of the channel with the *average* information content of the messages it carries. This notion is not one that appears in Shannon's paper, and plays no role in his results. However, we may explore the notion through an example.

We suppose that the base for the logarithms is 2, so that both the above quantities may be expressed in bits. We suppose further that we are concerned with sequences of independent binary messages – $n_m = 2$ and each individual message is either m_1 or m_2.

If $p_1 = p_2 = 0.5$, we have $H = 1$ bit. That is, each message carries (in Shannon terms) one bit of information. A sequence of n independent messages like this would carry n bits. This is all as we would expect. However, if (say) $p_1 = 0.89$ and $p_2 = 0.11$, H is considerably reduced, to about half a bit.

What does the statement $H = 0.5$ mean? The answer lies in Shannon's theoretical results. We can express one primary result, as applied to this example, as follows. Given a long sequence of n independent messages as above, with $p_1 = 0.89$ and $p_2 = 0.11$, it is possible to recode them into approximately $n/2$ binary messages without loss of information (so that the recipient could exactly recover the original sequence). In other words, *because of the imbalance of the probabilities*, we could compress the data by a factor of approximately 2.

This is a very powerful result. The theory does not tell us how to do the recoding – various methods exist, approaching the optimum to different degrees – but merely that it is possible. Note also that the result has little to do with the individual probabilities – it has to do with the imbalance between them.

But now let us consider a pointwise measure such as Equation 4. If we assume that this measure does represent the amount of information in a message, then the receipt of message m_1 will yield $I_1 = 0.17$ bits, while m_2 will yield $I_2 = 3.18$ bits. The question is, what does this mean? Because the I measure plays no role in the theory as originally proposed by Shannon, he gives us no guidance here. Nor am I aware of any other result in later developments, that might give I the same kind of status that is enjoyed by H.

In fact, it can be argued that this interpretation is at odds with Shannon theory. In Shannon theory, we start by waiting for a message, with certain probabilistic expectations. Then we receive it, and therefore discover what it is. It is the transformation from not knowing to knowing that conveys (in this case) the half-bit of information. By contrast, in the pointwise view, the amount of information depends on exactly what the message is.

One writer on information theory (Gray, 1990) has the following remark:

> Information measures are important because coding theorems exist imbuing them with operational significance and not because of intuitively pleasing aspects of their definitions.

This seems exactly right. Equation 3 is a good definition of an information measure precisely because it has operational significance. It is not at all clear that the same can be said of Equation 4.