Relevance Information: A Loss of Entropy but a Gain for IDF?

Arjen P. de Vries
CWI
PO Box 94079
1090 GB Amsterdam
The Netherlands
arjen@acm.org

Thomas Roelleke
Queen Mary University of London
Mile End Road
London, E1 4NS
United Kingdom
thor@dcs.qmul.ac.uk

ABSTRACT

When investigating alternative estimates for term discriminativeness, we discovered that relevance information and *idf* are much closer related than formulated in classical literature. Therefore, we revisited the justification of *idf* as it follows from the binary independent retrieval (BIR) model. The main result is a formal framework uncovering the close relationship of a generalised *idf* and the BIR model. The framework makes explicit how to incorporate relevance information into any retrieval function that involves an *idf*-component.

In addition to the *idf*-based formulation of the BIR model, we propose Poisson-based estimates as an alternative to the classical estimates, this being motivated by the superiority of Poisson-based estimates for the within-document term frequencies. The main experimental finding is that a Poisson-based *idf* is superior to the classical *idf*, where the superiority is particularly evident for long queries.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Theory, Experimentation.

Keywords: Information retrieval, relevance feedback, probability of relevance, binary independent retrieval model, inverse document frequency.

1. INTRODUCTION

Virtually all information retrieval systems estimate the relevance of documents to a query by comparing term statistics for the document at hand to those of the collection and, if available, a set of known relevant and/or non-relevant documents. Combining within-document term frequency (referred to as *tf*) with the inverse document frequency (referred to as *idf*) has consistently been proven successful on a variety of retrieval benchmarks.

The intuition underlying this combination of *tf* and *idf* is that frequent occurrence of query terms in a document makes the document more likely to be relevant to the query, but only if these terms are also discriminative. A term is discriminative if it does not occur in many different documents, or, in other words, when its inverse document frequency is high. Many publications point out how a $tf \cdot idf$ measure results naturally from either purely probabilistic or information theoretic arguments, and although they differ (sometimes quite significantly) in the details, in the end, all of these authors conclude that the theoretical argument presented fit nicely with the well-known experimental success.

Our investigation considers the role of relevance information in retrieval systems based on $tf \cdot idf$ ranking functions. Exploiting relevance information to revise the importance weights of given query terms and/or to expand the query with new terms usually improves the retrieval quality significantly, i.e., leads to a better ranking of documents than the one obtained by the initial provided by the user.

The main goal of this paper is to investigate how the estimate of term discriminativeness, represented by the term's *idf*, should be updated after relevance information becomes available. From an information theoretic argument, adding relevance information to a retrieval system corresponds to a loss of entropy. Without relevance information, the corpus of all documents is the foundation for estimating the discriminativeness (informativeness) of each term. With relevance information, we should remove the documents for which relevance information is available (let it be positive or negative relevance feedback) from the corpus of all documents.

We view *idf* primarily as a *summary statistic* of *a* set of documents. Church and Gale have demonstrated that *idf* is an attractive summary statistic of term sets, because it is more stable across collections of different years and different sources than alternatives such as, e.g., the variance of term occurrence in documents [2].

The classical binary independent retrieval (BIR) model [8] explains *idf* as the ranking that results from the situation where we have no relevance information. We have found that considering *idf* as a statistic of the sets of relevant and non-relevant documents provides an even stronger connection to the BIR model. We observed that the relevance-based term weights in the weighting schemes known as F1–F4 can be expressed conveniently as summations of the *idf* values of the different sets involved.

An open question remains how the discriminativeness of

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

a term should be measured. The classical *idf* is based on the occurrence probability $P(t|c) = n_D(t,c)/N_D(c)$, where $n_D(t,c)$ is the number of documents in collection c in which term t occurs and N_D is the number of documents in the collection. For the within-document term frequency, it is known that alternative estimates, such as those based on 'lifting' [13] and those based on Poisson approximations [6, 7], lead to improved retrieval results. The question that follows directly is whether the arguments motivating these alternative estimations apply equally to the estimate used for term discriminativeness.

The paper is structured as follows. Section 2 gives a concise overview of related work. Section 3 presents the main results of our research, where we relate the *idf* summary statistic over sets of relevant and non-relevant documents to the term weighting schemes of the Binary Independent Retrieval (BIR) model. Section 4 discusses the impact of our results for retrieval systems using *tf·idf* ranking, and Section 5 presents an experimental evaluation. We summarise the contributions of this work in the final Section 6.

2. BACKGROUND

Robertson recently surveyed research attempting to find plausible theoretical models explaining why *tf· idf* is a good approach to ranking [9]. The same work also presents the BM25 ranking function as another instantiation of *tf· idf* for ranked retrieval. Although we did not set out to provide a theoretical explanation for *tf · idf* approaches, the results obtained are related to these works, especially where we contribute into revealing a closer relationship of *idf* measures to the Binary Independent Retrieval (BIR) model [8].

From the start, our primary goal has been to investigate the role of the *idf* summary statistic in the incorporation of relevance information. Ruthven and Lalmas' review of approaches to relevance feedback [12] pointed us into the direction of updating *idf* according to the classic probabilistic model. Like [2], we approach *idf* as a robust statistic of term occurrence that helps identify 'good' keywords. Other related work in information retrieval theory includes the discussion of disjoint term spaces in [15], and in particular their proposal to represent discriminativeness probabilistically.

Our research can also be viewed as an extension of [10], where inverse document frequency has been related to the probability of being informative. Especially the results obtained on the TREC collections presented in Section 5 can be considered an experimental investigation into the applicability of the idea proposed there to define the probability of a term being informative based on the assumption of documents as independent rather than disjoint events, which motivates the use of Poisson probabilities.

3. IDF AND THE PROBABILITY OF RELEVANCE

This section discusses the relationship between the IDF and the probability of relevance. We review briefly the *idf* definition (section 3.1), and the probability of relevance (section 3.2), and the BIR model (section 3.3). Then, section 3.4 presents the *idf*-based formulation of the BIR model. Section 3.5 gives an intuitive explanation, based on the idea of 'virtual documents', for the parameter smoothing applied for dealing with singularities of the BIR model. Finally, we propose and investigate in section 3.6 Poisson-based estimates for probabilities such as $P(t|r)$ and $P(t|c)$ (probability that t occurs in relevant documents and probability that t occurs in the documents of the collection).

3.1 IDF

The *idf* is defined as the logarithm of the probability that a term occurs in the collection. Let c be the collection, let t be a term, let $N_D(c)$ be the number of documents in c, and let $n_D(t,c)$ be the number of documents in c in which t occurs. Then, the *idf* is defined as follows:

$$\begin{aligned} P(t|c) &:= \frac{n_D(t,c)}{N_D(c)} \\ idf(t,c) &:= -\log P(t|c) \end{aligned}$$

It reflects the discriminativeness of term t in collection c.

3.2 Probability of Relevance

The probability of relevance is denoted as $P(r|d,q)$ where r is the relevance event, d is the document event, and q is the query event.[1]

Bayes' theorem gives:

$$P(r|d,q) = \frac{P(r,d,q)}{P(d,q)}.$$

Next, we split the conjunction r, d, q. Once we let d depend on q, and once we let q depend on d.

$$\begin{aligned} P(r|d,q) &= \frac{P(d|q,r) \cdot P(r|q) \cdot P(q)}{P(d,q)} \quad (1) \\ &= \frac{P(q|d,r) \cdot P(r|d) \cdot P(d)}{P(d,q)} \quad (2) \end{aligned}$$

Equation 1 is the foundation of the binary independent retrieval (BIR) model and equation 2 is the foundation of the language modelling (LM) approaches.

Using the odds $O(r|d,q) = P(r|d,q)/P(\bar{r}|d,q)$ instead of $P(r|d,q)$ leads to the same ranking, but $P(d,q)$, $P(d)$ and $P(q)$ drop out.

$$\begin{aligned} O(r|d,q) = \frac{P(r|d,q)}{P(\bar{r}|d,q)} &= \frac{P(d|q,r) \cdot P(r|q)}{P(d|q,\bar{r}) \cdot P(\bar{r}|q)} \\ &= \frac{P(q|d,r) \cdot P(r|d)}{P(q|d,\bar{r}) \cdot P(\bar{r}|d)} \end{aligned}$$

This elegant formulation of the foundation and difference of the BIR model and LM approaches can be found in [5].

Now, to instantiate the model, documents and queries are considered as conjunctions of features, denoted with x_t. If we assume the features to be independent, we obtain:

$$\begin{aligned} P(d|q,r) &= \prod_t P(x_t|q,r) \\ P(q|d,r) &= \prod_t P(x_t|d,r) \end{aligned}$$

Usually, the words (terms) occurring in documents and queries are considered as features.

3.3 Binary Independent Retrieval (BIR) Model

The BIR model estimates the probability $P(d|q,r)$ based on the presence and absence of terms. The probability

[1] Note that $P(R|d,q)$ is also commonly used in literature. We prefer however a consistent use of lower case to denote events (as opposed to random variables).

$P(X_t = 1|q,r)$ is the probability that the term t occurs in relevant documents. Analogously, $P(X_t = 0|q,r)$ is the probability of its absence.

As usual in formulations of the BIR model, we define abbreviations for those probabilities. We set $a_t := P(X_t = 1|q,r)$ and $b_t := P(X_t = 1|q,\bar{r})$. (We deviate from the more common choice of p_t and q_t because this frequently causes confusion with probabilities and queries).

Since X_t is a binary random variable, we can rewrite the probability for an event x_t as follows [14]:

$$\begin{aligned} P(x_t|q,r) &= a_t^{x_t} \cdot (1-a_t)^{1-x_t} \\ P(x_t|q,\bar{r}) &= b_t^{x_t} \cdot (1-b_t)^{1-x_t} \end{aligned}$$

We obtain:

$$\begin{aligned} O(r|d,q) &= \frac{P(d|q,r)}{P(d|q,\bar{r})} \cdot O(r|q) \\ &= \prod_t \left(\frac{P(x_t|q,r)}{P(x_t|q,\bar{r})} \right) \cdot O(r|q) \\ &= \prod_t \left(\frac{a_t^{x_t} \cdot (1-a_t)^{1-x_t}}{b_t^{x_t} \cdot (1-b_t)^{1-x_t}} \right) \cdot O(r|q) \end{aligned}$$

The log transformation yields:

$$\begin{aligned} &\log O(r|d,q) = \\ &= \sum_t \log \frac{a_t^{x_t} \cdot (1-a_t)^{1-x_t}}{b_t^{x_t} \cdot (1-b_t)^{1-x_t}} + \log O(r|q) \\ &= \sum_t \log \frac{a_t^{x_t} \cdot (1-b_t)^{x_t} \cdot (1-a_t)}{b_t^{x_t} \cdot (1-a_t)^{x_t} \cdot (1-b_t)} + \log O(r|q) \\ &= \sum_t \left(x_t \cdot \log \frac{a_t \cdot (1-b_t)}{b_t \cdot (1-a_t)} \right) + \\ &\qquad \sum_t \log \frac{1-a_t}{1-b_t} + \log O(r|q) \\ &= \sum_t (x_t \cdot h_t) + Q \end{aligned}$$

Q is a constant depending only on the query. Since it has no effect on the ranking of the documents, we do not consider it further.

Term weight h_t, the *term discriminativeness* of t, equals:

$$h_t := \log \frac{a_t \cdot (1-b_t)}{b_t \cdot (1-a_t)}$$

We write $h(t,c)$ to refer to the discriminativeness of term t in collection c.

Estimation of probabilities a_t and b_t uses the proportions of relevant and non-relevant documents in the collection. First, we introduce the classical and our alternative notation (also used in [11]). As the reader will notice, the classical formulation makes use of r, which also denotes the relevance event introduced before. We apologise for this overloading, but apply the classical notation where it improves the readability.

r_t	$n_D(t,r)$	number of relevant documents with t
R	$N_D(r)$	number of relevant documents
n_t	$n_D(t,c)$	number of documents with t
N	$N_D(c)$	number of documents

Review that, in our notation (given in the second column), r and c are sets of documents: r is the set of relevant documents and c is the set of all documents (the collection). This notation is better suited to clarify dualities between estimates involving all documents, the relevant documents, or the non-relevant documents, and the document or (within-document) term frequencies.

Next, consider the four alternatives for estimation, classically denoted by F1 to F4 [8].

F1: $h_t := \log \frac{r_t/R}{n_t/N}$: term occurrence is independent in *all* documents; consider occurrence only, not absence

F2: $h_t := \log \frac{r_t/R}{(n_t-r_t)/(N-R)}$: term occurrence is independent in *the non-relevant* documents; consider occurrence only, not absence

F3: $h_t := \log \frac{r_t/R\cdot(1-n_t/N)}{n_t/N\cdot(1-r_t/R)}$: term occurrence is independent in *all* documents; consider occurrence and absence

F4: $h_t := \log \frac{r_t/R\cdot(1-(n_t-r_t)/(N-R))}{(n_t-r_t)/(N-R)\cdot(1-r_t/R)} = \log \frac{r_t\cdot((N-R)-(n_t-r_t))}{(n_t-r_t)\cdot(R-r_t)}$: term occurrence is independent in *the non-relevant* documents; consider occurrence and absence

3.4 IDF and the BIR Model

From the formulation in the previous section, the term weights h_t can be viewed as sums of *idf*-values.

Reconsider the definition of *idf*:

$$\mathit{idf}(t,c) := -\log \frac{n_D(t,c)}{N_D(c)} = -\log \frac{n_t}{N}$$

Using the set relevant documents instead of the full collection, we obtain:

$$\mathit{idf}(t,r) := -\log \frac{n_D(t,r)}{N_D(r)} = -\log \frac{r_t}{R}$$

This dual statistic reflects the discriminativeness of term t in the set of relevant documents r.

Consider the sum of *idf*-values we obtain for each of the four settings of h_t:

F1: $h_t = -\mathit{idf}(t,r) + \mathit{idf}(t,c)$: Here, the BIR model can be explained as follows: Given relevance information, the *idf* of the terms based on the relevant documents is subtracted from the standard *idf*. That means: If an overall rare term occurs rarely among the relevant documents, then $\mathit{idf}(t,r)$ is large, thus, h_t is significantly smaller than $\mathit{idf}(t,c)$. It is smaller, since, although the term is discriminative overall, for this particular query, the term is not leading to relevant documents. If an overall rare term is frequent among the relevant documents, then $\mathit{idf}(t,r)$ is small, and we obtain $h_t \approx \mathit{idf}(t,c)$. We obtain $h_t = \mathit{idf}(t,c)$ if t occurs in all relevant documents. Without relevance information, each term occurs in all relevant documents, because the set of relevant documents is empty.

F2: $h_t = -\mathit{idf}(t,r) + \mathit{idf}(t,\bar{r})$: Here, $\bar{r}$ denotes the set of non-relevant documents, $\bar{r} = c \setminus r$ (also known as the 'complement method'). Positive and negative relevance feedback is exploited in this second possible setting of h_t. The discriminativeness of a term is now based on the set of non-relevant documents, rather than on the set of all relevant as it was the case for the first setting.

F3: $h_t = -\mathit{idf}(t,r) - \mathit{idf}(\bar{t},c) + \mathit{idf}(\bar{t},r) + \mathit{idf}(t,c)$: The third setting shows all four combinations of term occurrence or absence, and relevant or all documents.

$-\mathit{idf}(t,r)$ is the discriminativeness of t in relevant documents. Reduces h_t only slightly for terms frequent in r, but strongly for non-frequent terms.

$-\mathit{idf}(\bar{t},c)$ reduces h_t only slightly if the event "t is absent in the collection" is frequent. Then, t occurs rarely, i.e., $\mathit{idf}(t,c)$ is large.

$\mathit{idf}(\bar{t},r)$ is the discriminativeness of the absence of t in r. If the event "t is absent in relevant document" is rare, then t occurs frequently in r, and this leads to a high h_t. Thus, $\mathit{idf}(\bar{t},r)$ supports the message of $-\mathit{idf}(t,r)$, i.e., rare absence in relevant documents is good, means frequent occurrence in relevant document is good.

F4: $h_t = -\mathit{idf}(t,r) - \mathit{idf}(\bar{t},\bar{r}) + \mathit{idf}(t,\bar{r}) + \mathit{idf}(\bar{t},r)$: The fourth setting shows all possible four combinations of term occurrence or absence, and relevant or non-relevant documents.

$-\mathit{idf}(t,r)$ and $\mathit{idf}(\bar{t},\bar{r})$ have been discussed for the previous cases.

$-\mathit{idf}(\bar{t},\bar{r})$ supports $-\mathit{idf}(t,\bar{r})$, i.e., if the event "t is absent in non-relevant document" is frequent (means, t hardly occurs in non-relevant documents), $-\mathit{idf}(\bar{t},\bar{r})$ diminishes, i.e., it leads to little reduction for h_t.

$\mathit{idf}(t,\bar{r})$ is the discriminativeness of t in non-relevant documents. Without relevance information, $\mathit{idf}(t,\bar{r}) = \mathit{idf}(t,c)$.

This representation of the term weights of the BIR model based on a sum of *idf*-values gives a clear justification of the *idf* through the BIR model. It generalises the results and claim in [9], page 512: "It is now apparent that we can regard IDF as a simple version of the RSJ weight applicable when we have no relevance information."

When no relevance information is available (maximum entropy), the classical *idf* corresponds to the measure of discriminativeness. The more relevance information becomes available, the more the term's discriminativeness (in the *remainder* of the collection) will decrease (a loss of entropy). It decreases because the corpus of documents for which no relevance information is available is smaller, and, more importantly, because the discriminativeness of the query term in the relevant documents is *subtracted* from the discriminativeness obtained for all (or all non-relevant) documents.

3.5 Dealing with the Singularities

The estimates for the discriminativeness weights h_t derived above have to be adapted for dealing with the singularities. The following cases require adaptation of the estimates:

- $R = N_D(r) = 0$: No relevance information given.
- $r_t = n_D(t,r) = 0$: Term t does not occur in any relevant document. Then, $\log P(t|r)$ is not defined.
- $N = N_D(c) = 0$: Empty collection.
- $n_t = n_D(t,c) = 0$: Term t does not occur in any document.
- $R = r_t = N_D(r) = n_D(t,r)$: Term t occurs in all relevant documents. Singularity for F3 and F4.
- $N = n_t = N_D(c) = n_D(t,c)$: Term t occurs in all documents. Singularity for F3 and F4.
- $n_t = r_t = n_D(t,c) = n_D(t,r)$: All documents in which term t occurs are relevant. Singularity for F4.
- $N - R = n_r - r_t = N_D(c) - N_D(r) = n_D(t,c) - n_D(t,r)$: Term t occurs in all non-relevant documents. Singularity for F4.

A common reformulation (smoothing) of the estimates is to add constants to the expressions in numerator and denominator.

The classic smoothing proposed for F4 (see [9]) adds a constant value 0.5 to each document count:

$$h_t := \log \frac{(r_t + 0.5) \cdot ((N - R) - (n_t - r_t) + 0.5)}{(R - r_t + 0.5) \cdot (n_t - r_t + 0.5)}$$

Although the literature has given various mathematical motivations for this constant, we like to offer a particularly intuitive explanation. Consider that we add two virtual documents to each collection, one of which is relevant. Assuming that each term occurs in half of the virtual and relevant documents gives the following probability estimates:

$$P(t|c) := \frac{n_D(t,c) + 1}{N_D(c) + 2} \quad P(t|r) := \frac{n_D(t,r) + 0.5}{N_D(r) + 1}$$

Based on these adapted estimates, we obtain the smoothed F4:

$$\begin{aligned} h_t = & \log \frac{(r_t{+}0.5)/(R{+}1) \cdot (1 - (n_t - r_t{+}0.5)/(N{-}R{+}1))}{(1 - (r_t{+}0.5)/(R{+}1)) \cdot (n_t - r_t{+}0.5)/(N{-}R{+}1)} \\ = & \log \frac{(r_t + 0.5) \cdot ((N - R) - (n_t - r_t) + 0.5)}{(R - r_t + 0.5) \cdot (n_t - r_t + 0.5)} \end{aligned}$$

So, extending the document sets with two virtual documents is sufficient to explain the classical parameter adaptation. Next, consider four virtual documents; two of which are relevant, all terms occur in half of the documents, and all terms occur in half of the relevant documents.

$$P(t|c) := \frac{n_D(t,c) + 2}{N_D(c) + 4} \quad P(t|r) := \frac{n_D(t,r) + 1}{N_D(r) + 2}$$

This solution feels more comfortable, because all numbers are integers; they can be interpreted as document counts. We obtain:

$$\begin{aligned} h_t = & \log \frac{(r_t + 1)/(R + 2) \cdot (1 - (n_t - r_t + 1)/(N - R + 2))}{(1 - (r_t + 1)/(R + 2)) \cdot (n_t - r_t + 1)/(N - R + 2)} \\ = & \log \frac{(r_t + 1) \cdot ((N - R) - (n_t - r_t) + 1)}{(R - r_t + 1) \cdot (n_t - r_t + 1)} \end{aligned}$$

As a general explanation, we derive:

$$P(t|c) := \frac{n_D(t,c) + 2\epsilon}{N_D(c) + 4\epsilon} \quad P(t|r) := \frac{n_D(t,r) + 1\epsilon}{N_D(r) + 2\epsilon}$$

The number of virtual documents that is added to the collection reflects the uncertainty about relevance. We assume that each term t occurs in half of the virtual documents (since occurrence is a binary event), that half of the virtual documents are relevant (since relevance is a binary event), and that each term t occurs in half of the relevant documents. As demonstrated, these assumptions explain the classical smoothing proposed for F4 by defining $\epsilon = 0.5$.

3.6 Poisson-based Probability Estimation

Probability estimation for $P(t|c)$ and $P(t|r)$ and related probabilities have so far been estimated on what we refer to as n/N estimates and their adaptations. Consider a set of N trials, and let n_t be the number of trials in which the event t is true. Then, $P(t) = n_t/N$ is the probability that the event is true among N trials. We refer to this estimate as the disjointness-based estimate, since it can be explained by the theorem of total probability as $P(t) := \sum_d P(t|d) \cdot P(d)$, where the events d are disjoint and exhaustive.

The appropriateness of these estimates can however be questioned, in particular in the light of the summation of *idf*-values as follows from the BIR model. For, the sets involved (set of relevant documents, set of non-relevant documents, set of all documents) are different in cardinality: usually, only a small fraction of all documents forms the set of relevant documents.

Compare this to estimates for $P(t|d)$, i.e., the probability that a term describes (occurs in) a document. While the disjointness-based probability corresponds to the probability that t occurs, the Poisson-based probability of term occurrence is the probability that t occurs at least once in n_t trials. The usefulness of a Poisson-based probability is well-known for the within-document term frequency (referred to as *tf* or *lf*) of a term t in a document d [6, 7].

Let $n_L(t, d)$ be the number of locations at which t occurs in d. Then, the Poisson-based estimate of term (location) frequency is defined as follows:

$$P(t|d) := \mathit{tf}(t, d) = \mathit{lf}(t, d) = \frac{n_L(t, d)}{K_d + n_L(t, d)}$$

Here, $P(t|d)$ is the probability that term t is representative for document d. The steep rise of the within-document term frequency for small occurrences ($n_L(t, d)$ relatively small) leads to a superior retrieval quality. For example, let K be the average occurrence frequency ($K := 1/N_T \cdot \sum_t n_t$, where N_T is the number of events/terms). We obtain $P(t) = 0.5$ if n_t is the average occurrence, $P(t) < 0.5$ for n_t less than the average occurrence, and $P(t) > 0.5$ for n_t greater than the average occurrence.

This effect of a steep rise is also reflected in the well-known lifting function with lifting (zooming) factor z_r ([13] and related publications):

$$\mathit{tf}(t, d) = z_d + (1 - z_d) \cdot \frac{n_L(t, d)}{N_L(d)}$$

For $z_d := 0.5$, the lifting yields $\mathit{tf}(t, d) \approx 0.5$, if $n_L(t, d) = 1$. Compare this to the Poisson approximation, where for $K_d = 1$, we obtain $\mathit{tf}(t, d) = 0.5$, if $n_L(t, d) = 1$.

Looking at the sum of *idf*-values as presented in section 3.3, we deal with sets of significantly different cardinality. The set of relevant documents is much smaller than the set of non-relevant and the set of all documents. This different cardinality and the success of the lifting of the probability $P(t|d)$ are the motivations for investigating Poisson-based estimates for $P(t|c)$ and $P(t|r)$.

Review the Poisson-based probability estimation for $P(t|c)$, and $P(t|r)$:

$$\begin{aligned} P(t|c) &:= \mathit{df}(t, c) := \frac{n_D(t, c)}{K_c + n_D(t, c)} \\ P(t|r) &:= \mathit{df}(t, r) := \frac{n_D(t, r)}{K_r + n_D(t, r)} \end{aligned}$$

Compare these also to the formulation of $P(t|d)$ given before. The notation shows the strong analogy of the probabilities involved.

Section 5 reports the experimental results obtained for Poisson-based estimates. Before we started the experimental investigation, we looked at an analytical experiment where we investigate the nature of the Poisson-based estimate. Consider the analytical experiment in Table 1. Let a collection with $N_D(\bar{r}) = 10^6$, one million documents, be given, and let a query with $N_D(r) = 10$, i.e., ten relevant documents, be given. For ease of reading, we use logarithm base 10 for the numerical illustration (the base of the logarithm does not matter for ranking purpose, as the base is a constant, term-independent factor).

The table shows some numerical values for the discriminativeness $h_t = \mathit{idf}(t, \bar{r}) - \mathit{idf}(t, r)$, earlier also referred to as the F2 estimate. The table shows six cases of possible term distributions, and the cases (a), (d), (e) and (f) (bold face) illustrate the difference between the $n(t, x)/N$ and the $n(t, x)/(K_x + n(t, x))$ estimates. Here, x represents a set of documents: the set r of relevant documents, or the set $\bar{r}$ of non-relevant documents. The main findings from this analytical investigation are:

1. The classical estimate n_t/N leaves h_t to be high for terms rare in the collection; the effect of the occurrence of the term in the relevant document is minor (see (a), (b), and (c)). h_t is mainly determined by $\mathit{idf}(t, \bar{r})$.

2. The Poisson-based estimate leads to a dramatic impact on h_t for terms that occur rarely in the relevant documents (see (a) and (d)). For a rare term, the sum of idf_P-values is zero (case (a))! For a more frequent term, $\mathit{idf}_P(t, \bar{r})$ remains the main impact of t on the RSV (cases (b) and (c)).

3. The classical estimate n_t/N leaves $\mathit{idf}(t, \bar{r})$ to be the dominant factor even for terms that are relatively frequent in $\bar{r}$, and the frequency of t in r has hardly an effect (see (d), (e), and (f)).

4. The Poisson-based estimate leads to small $\mathit{idf}_P(t, \bar{r})$ values for terms that are relatively frequent. The impact of $\mathit{idf}_P(t, r)$ is strong (see (d)) for terms rare among the relevant. For a term relatively frequent in relevant and non-relevant documents, the effect on the RSV is minimal (see (e) and (f)). As the analytical investigation shows, this cancelling out of terms is already happening for terms that are little to medium frequent, if the parameter K of the Poisson-approximations is small (for the analytical experiment, $K_{\bar{r}} = K_r = 1$).

This analytical experiment illustrates the different nature of the classical n_t/N (*idf*) estimate and the Poisson-based estimate $n/(K{+}n)$ (idf_P). As our TREC experiments show, the setting of K has a major impact on the retrieval quality. Before we report on the experiments, we first summarise the impact of our results on $\mathit{tf} \cdot \mathit{idf}$-based retrieval functions.

4. IMPACT ON TF-IDF-BASED RANKING

One result of our investigation is that any retrieval system based on $\mathit{tf} \cdot \mathit{idf}$ ranking may incorporate relevance information by replacing its *idf*-component by h_t, where h_t is a

Table 1: Analytical Experiment: Comparison of h_t for $n(t,x)/N$ and $n(t,x)/(K_x + n(t,x))$

			$n(t,x)/N$			$n(t,x)/(1+n(t,x))$		
	$n_D(t,r)$	$n_D(t,\bar{r})$	$\mathit{idf}(t,r)$	$\mathit{idf}(t,\bar{r})$	$h_t =$ $-\ \mathit{idf}(t,r) + \mathit{idf}(t,\bar{r})$	$\mathit{idf}_P(t,r)$	$\mathit{idf}_P(t,\bar{r})$	$h_t =$ $-\ \mathit{idf}_P(t,r) + \mathit{idf}_P(t,\bar{r})$
a	1	1	$-\log\frac{1}{10}$	$-\log\frac{1}{10^6}$	$-1+6=5 \approx \mathit{idf}(t,\bar{r})$	$-\log\frac{1}{2}$	$-\log\frac{1}{2}$	0
b	5	1	$-\log\frac{5}{10}$	$-\log\frac{1}{10^6}$	$>5 \approx \mathit{idf}(t,\bar{r})$	$-\log\frac{5}{6}$	$-\log\frac{1}{2}$	$\approx \mathit{idf}_P(t,\bar{r})$
c	10	1	$-\log\frac{10}{10}$	$-\log\frac{1}{10^6}$	$6 = \mathit{idf}(t,\bar{r})$	$-\log\frac{10}{11}$	$-\log\frac{1}{2}$	$\approx \mathit{idf}_P(t,\bar{r})$
d	1	100	$-\log\frac{1}{10}$	$-\log\frac{100}{10^6}$	$-1+4=3 \approx \mathit{idf}(t,\bar{r})$	$-\log\frac{1}{2}$	$-\log\frac{100}{101}$	$\approx -\mathit{idf}_P(t,r)$
e	5	100	$-\log\frac{5}{10}$	$-\log\frac{100}{10^6}$	$>3 \approx \mathit{idf}(t,\bar{r})$	$-\log\frac{5}{6}$	$-\log\frac{100}{101}$	≈ 0
f	10	100	$-\log\frac{10}{10}$	$-\log\frac{100}{10^6}$	$4 = \mathit{idf}(t,\bar{r})$	$-\log\frac{10}{11}$	$-\log\frac{100}{101}$	≈ 0

sum of *idf*-values derived from the sets of relevant and non-relevant documents as presented in section 3.4. This result coincides with the BM25 ranking formula, and extends the relationship of *idf* and BIR model as presented in [9].

Consider the definition of a *tf*·*idf*-based retrieval function, and we illustrate the replacement of *idf* by the discriminativeness measure h_t:

$$\begin{aligned} RSV_{tf-idf} &:= \sum_t \mathit{idf}(t,c) \cdot \mathit{tf}(t,d) \cdot \mathit{tf}(t,q) \\ &= \sum_t h(t,c) \cdot \mathit{tf}(t,d) \cdot \mathit{tf}(t,q) \end{aligned}$$

The probabilities based on the within-document and within-query term frequencies are estimated as follows:

$$\begin{aligned} P(t|d) := \mathit{tf}(t,d) &:= \frac{n_L(t,d)}{K_d + n_L(t,d)} \\ P(t|q) := \mathit{tf}(t,q) &:= \frac{n_L(t,q)}{K_q + n_L(t,q)} \end{aligned}$$

The factors K_d and K_q serve for normalisation purposes. The BM25 (as the currently most successful *tf* · *idf*-based retrieval function) proposes $h(t,c)$ in place of *idf*, and adds various constants to the retrieval function. The motivation for the constants is to leverage the effect of $h(t,c)$, $\mathit{tf}(t,d)$, and $\mathit{tf}(t,q)$, and to fit the retrieval function to the nature of the collection for which it is applied.

To illustrate, we substitute $h(t,c)$ with the *idf*-based formulation of F1:

$$RSV_{tf-idf} = \sum_t \left[-\mathit{idf}(t,r) + \mathit{idf}(t,c)\right] \cdot \mathit{tf}(t,d) \cdot \mathit{tf}(t,q)$$

The factor $\mathit{idf}(t,c)$ is the classical *idf*. The factor $-\mathit{idf}(t,r)$ represents the relevance information. This factor is negative, since $h(t,c)$ is smaller than $\mathit{idf}(t,c)$ for terms that occur rarely in relevant documents. The example shows that for $h(t,c)$ based on F1, $RSV_{tf \cdot idf}$ corresponds to a classic *tf* · *idf* if no relevance information is available, and, the formulation also shows that with $\mathit{idf}(t,c)$ only we actually assume that t occurs in all relevant documents ($\mathit{idf}(t,r) = 0$).

This framework makes the relationship between *idf* and relevance information fully explicit.

5. EXPERIMENTS

So far, the results of this paper have been of theoretical nature: relevance feedback leads to a revised *idf* in *tf* · *idf*-based retrieval functions. In the derivation of this result, we proposed Poisson-based probability estimates instead of the classic (n/N-based) ones. As the weighting schemes corresponding to the BIR model combine *idf*-values of sets with highly different cardinalities, we investigate experimentally whether a Poisson-based estimation makes sense for probabilities $P(t|c)$ and $P(t|r)$.

5.1 Experimental Setup

The experimental evaluation consists of a first series of *adhoc* retrieval experiments to investigate the effect of Poisson-based estimates for *idf* in a setting in absence of relevance information, and a subsequent series of *routing* experiments to investigate its effect on processing relevance information using the four variants of the BIR model to exploit relevance information.

We have used the TREC test collections developed at TREC-7 and TREC-8 (using the data of 'disks 4 and 5', and topic sets 351–400 and 401–450 respectively). Preprocessing of the documents included neither stop-word removal nor stemming, resulting in the same experimental setup as the one described in [4]. However, terms in the query that occur on Van Rijsbergen's stop-word list have been removed from the query text.

Like the adhoc experiment, the routing experiment follows the setup described in [4], where the training data consists of the Los Angeles Times articles and the test data consists of the remaining data. This division of the TREC data in training and testing is a natural one for a routing task, because the training articles date from 1989 and 1990, in time preceding the rest of the collection which contains publications from 1991 until 1994.

The retrieval system has been developed by extending the open-source main memory database system MonetDB[2] with some minimal extra functionality to support information retrieval. To validate the basic workings of this retrieval system implementation, we first reproduced the baseline adhoc retrieval approach presented in [4], and obtained indeed the mean average precision (MAP) reported there.[3] We then proceeded to implement the BM25 ranking, where we choose $k_1 = 1.2$ and $k_3 = 1000$ following the Okapi TREC papers. We have set the length normalisation parameter b to 0.7627, as derived analytically in [1].

5.1.1 *Adhoc experiments*

Table 2 presents an overview of the results on the TREC-7 and TREC-8 adhoc test collections. First, compare the bottom row displaying the baseline results of a standard BM25 ranking using *idf* estimation, to the other rows, showing the results for BM25 using idf_P instead of *idf*. On the short

[2] See `http://monetdb.cwi.nl/`.

[3] Notice that these results are lower than those reported in [3], which we attribute to differences in pre-processing.

Table 2: Mean Average Precision of adhoc retrieval experiments for topics created from title (T), description (D) and narrative (N). The bottom line is the normal *idf* estimate; the others are Poisson-estimated *idf* with varying K, where $N = 528,024$, and $\widehat{n_t} = \frac{1}{N}\sum_t n_t \approx 260$.

	TREC–7			TREC–8		
K	T	TD	TDN	T	TD	TDN
1	0.143	0.146	0.153	0.161	0.160	0.168
$\widehat{n_t}$	0.144	0.149	0.161	0.164	0.166	0.184
$N/100$	0.153	0.164	0.193	0.171	0.194	0.214
$N/50$	0.157	0.168	0.198	0.175	0.198	0.218
$N/10$	**0.163**	**0.176**	**0.204**	0.185	0.209	**0.225**
$N/3$	0.161	0.175	0.195	0.188	**0.211**	0.223
$N/2$	0.159	0.173	0.192	0.188	0.211	0.221
N	0.157	0.170	0.187	**0.188**	0.210	0.218
-	0.138	0.148	0.154	0.183	0.194	0.194

Table 3: Mean Average Precision of adhoc retrieval experiments for topics created from title (T), description (D) and narrative (N), using *idf* only.

	T	TD	TDN
TREC-7	0.124	0.095	0.041
TREC-8	0.136	0.111	0.064

TREC-7 queries, all of the idf_P results outperform the baseline. If K is sufficiently high, the mean average precision of runs using idf_P is better than the baseline regardless the topic length; $K = N/10$ gives near-best results for all of the experimental conditions. From these experiments, it seems fair to conclude that it makes sense to apply the Poisson-estimation to the occurrence frequencies underlying the inverse document frequency as a model of the discriminative power of query terms.

Tables 3 and 4 present the results obtained using $idf(t, c)$ weighting *only*, using *idf* and idf_P respectively. Performance of the title-only queries is still reasonably good when compared to the results in Table 2. Using *idf* on its own degrades however for longer queries, whereas idf_P results are impressive regardless the topic length.

5.1.2 Routing experiments

The routing experiments use the relevance assessments on the LA Times articles in the estimation of $idf(t, r)$. The baseline results are those obtained on the test data (not including the training data) without using any relevance information. We then apply the four weighting schemes of the BIR model, comparing the effect of relevance information obtained with the standard *idf* to that using the Poisson-based idf_P.

Table 4: Mean Average Precision of adhoc retrieval experiments for topics created from title (T), description (D) and narrative (N), using idf_P only ($K = N/10$).

	T	TD	TDN
TREC-7	0.133	0.143	0.127
TREC-8	0.136	0.158	0.137

Table 5: Mean Average Precision of routing experiments with weighting strategies, for topics created from title (T), description (D) and narrative (N). The Poisson-based *idf*, denoted idf_P, is parameterised with $K = N/10$.

		TREC–7			TREC–8		
Method		T	TD	TDN	T	TD	TDN
-	*idf*	0.125	0.135	**0.147**	**0.181**	**0.193**	0.194
F1	*idf*	**0.135**	**0.138**	0.144	0.154	0.180	**0.201**
F2	*idf*	**0.135**	**0.138**	0.144	0.154	0.180	**0.201**
F3	*idf*	0.135	0.137	0.139	0.153	0.177	0.195
F4	*idf*	0.135	0.137	0.139	0.153	0.177	0.195
-	idf_P	**0.161**	**0.178**	**0.206**	**0.181**	**0.213**	**0.237**
F1	idf_P	0.152	0.165	0.191	0.168	0.205	0.229
F2	idf_P	0.152	0.165	0.191	0.168	0.205	0.229
F3	idf_P	0.151	0.165	0.190	0.168	0.205	0.228
F4	idf_P	0.151	0.165	0.190	0.168	0.205	0.228

Table 6: Mean Average Precision of routing experiments with idf_P for $K = 1$, for weighting strategies, topics created from title (T), description (D) and narrative (N).

		TREC–7			TREC–8		
Method		T	TD	TDN	T	TD	TDN
-	idf_P	0.141	0.148	0.158	0.149	0.153	0.175
F1	idf_P	0.121	0.089	0.048	0.129	0.104	0.086
F2	idf_P	0.121	0.089	0.048	0.129	0.104	0.086
F3	idf_P	0.089	0.064	0.050	0.132	0.112	0.095
F4	idf_P	0.089	0.064	0.050	0.132	0.112	0.095

Table 5 summarises the routing results. Whereas with classical *idf*, relevance feedback obtained from the assessments on the Los Angeles Times articles improves results over the baseline, applying the BIR relevance weighting using the Poisson-based idf_P does not lead to results that improve upon *its* baseline. Notice however that *all* experimental results using idf_P result in higher mean average precision scores than all of those using *idf*, with or without relevance information. While for long queries, a marginal performance improvement is observed with standard *idf* weighting, the results after feedback remain significantly below the idf_P baseline. The differences in mean average precision between F1 and F2 seem negligible, as well as those between F3 and F4. Surprisingly, the weighting schemes based on term presence only (F1 and F2) outperform consistently those on term presence and absence (F3 and F4).

We presented only the experimental results for $K = N/10$, but the observations made hold for other choices of K as well. Table 6 shows an adverse effect of the Poisson weighting on the routing task performance (especially on the longer queries). Again, these are representative for all settings with small K.

5.2 Results and Analysis

The difference in performance between short and long queries in tables 3 and 4 can be partially explained by assuming that the user stating a short 'title' query selects very carefully the most discriminative terms with respect to relevance, while the description and narrative also contain noisy terms that do not directly relate to the underlying information need. The *idf* values do not differentiate between discriminative yet non-relevant terms and non-discriminative yet relevant terms in the query.

The Poisson-based *idf* yields the same order of terms with respect to their discriminativeness as the classical *idf* does. Therefore, the effect of using one or the other approach is stronger for long queries than for short queries.

Setting the K-parameters of the Poisson-estimates to small values corresponds to weakening the effect of *idf* on the RSV, whereas a large K strengthens the effect of *idf*. Consistently, we can observe that a strong *idf* leads to better retrieval results.

A Poisson-based *idf* with large K is consistently superior to the classical *idf*. This experimental result can be explained as follows from a theoretical point of view. The Poisson-based *idf* distinguishes stronger between discriminative terms, and distinguishes less between non-discriminative terms than the classical *idf*. As our results show, this nature of the Poisson-based *idf* of being stronger for rare terms and less strong for frequent terms leads to better retrieval quality. This result perfectly coincides with the superiority of the Poisson-based probability reflecting the within-document frequency.

Further in-depth analysis is needed to give an explanation as to why the ample training data in the routing experiment did not lead to a considerable improvement on the test data; though we like to point out that [4] also reports disappointing results with feedback in the routing experiment performed; so, maybe in the specific experimental setup chosen it is especially hard to gain improvements, and we should repeat the experiment on another (TREC) routing or filtering task.

6. SUMMARY AND CONCLUSIONS

This paper makes explicit the relationship between *idf* and the binary independent retrieval (BIR) model. The impact of this result is that relevance information can be incorporated into any $tf \cdot idf$-based retrieval function by revising the *idf* according to the relevance information available. Our result makes explicit that relevance information can be viewed as a loss of entropy. The discriminativeness of terms, initially high because of maximal entropy, decreases for two reasons. First, the event space for estimating term discriminativeness after feedback is smaller. Second, the discriminativeness of terms in relevant documents is subtracted from the original *idf*.

In addition to this theoretical framework for relating *idf* and the BIR model, and for justifying the incorporation of relevance information into a revised *idf*-component, we investigated Poisson-based probability estimations. Motivated by the success of Poisson-based estimates for within-document frequencies, we applied Poisson-estimates for the occurrence probabilities $P(t|r)$ (term occurrence in relevant documents) and $P(t|c)$ (term occurrence in all documents), which form the basis for the discriminativeness measures. The overall result is that the Poisson-based *idf* is superior to the classical *idf*, in particular for long queries. Unfortunately, we could not prove experimentally a positive effect in improving retrieval by relevance feedback.

Our next research steps include the transformation of *idf*-values for measuring discriminativeness (informativeness, respectively) into probabilities. This is a key issue for probabilistic retrieval models and probabilistic reasoning. For example, the framework described in [15] is based on a disjoint space of concepts, where we could apply an *idf*-based probability as an estimate of the probability of term informativeness. The refinement of those term space probabilities according to relevance feedback data has been an open problem. Though using *idf*-values as a basis for probability estimation poses a number of problems, the result of this paper on how to manage relevance information in an *idf*-based space might lead to new insights in how information theory and probability theory interrelate.

Acknowledgement

We thank Djoerd Hiemstra for providing pre-processed TREC data.

7. REFERENCES

[1] G. Amati. *Probability Models for Information Retrieval based on Divergence from Randomness.* PhD thesis, Glasgow University, June 2003.

[2] K.W. Church and W.A. Gale. Inverse document frequency: A measure of deviations from poisson. In *Third Workshop on Very Large Corpora, ACL Anthology*, 1995.

[3] A.P. de Vries and D. Hiemstra. The Mirror DBMS at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference TREC-8*, pages 725–734, Gaithersburg, Maryland, November 1999.

[4] D. Hiemstra, S.E. Robertson, and H. Zaragoza. Parsimonious Language Models for Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, July 2004.

[5] J. Lafferty and Ch. Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, chapter 1. Kluwer, 2002.

[6] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, London, et al., 1994. Springer-Verlag.

[7] S.E. Robertson, S. Walker, and M.M. Hancock-Beaulieu. Large test collection experiments on an operational interactive system: Okapi at TREC. *Information Processing and Management*, 31:345–360, 1995.

[8] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[9] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments. *Journal of Documentation*, 60(5):503–520, 2004.

[10] T. Roelleke. A frequency-based and a poisson-based probability of being informative. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada*, pages 227–234, 2003.

[11] T. Roelleke, T. Tsikrika, and G. Kazai. A general matrix framework for modelling information retrieval. *Journal on Information Processing & Management (IP&M)*, 2005. To appear.

[12] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[13] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information retrieval. In *11th International Conference on Research & Development in Information Retrieval*, pages 147–160, Grenoble, France, June 1988.

[14] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2. edition, 1979.

[15] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.