

Generalized Inverse Document Frequency

Donald Metzler
metzler@yahoo-inc.com
Yahoo! Research
2821 Mission College Blvd.
Santa Clara, CA 95054

ABSTRACT

Inverse document frequency (IDF) is one of the most useful and widely used concepts in information retrieval. There have been various attempts to provide theoretical justifications for IDF. One of the most appealing derivations follows from the Robertson-Sparck Jones relevance weight. However, this derivation, and others related to it, typically make a number of strong assumptions that are often glossed over. In this paper, we re-examine these assumptions from a Bayesian perspective, discuss possible alternatives, and derive a new, more generalized form of IDF that we call generalized inverse document frequency. In addition to providing theoretical insights into IDF, we also undertake a rigorous empirical evaluation that shows generalized IDF outperforms classical versions of IDF on a number of ad hoc retrieval tasks.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Theory

Keywords

inverse document frequency, formal models, estimation

1. INTRODUCTION

Inverse document frequency (IDF) is arguably one of the most important and widely used concepts in information retrieval. It was first introduced by Sparck Jones in 1972 [9] with the aim of improving automatic indexing and retrieval systems. Since that time, IDF has become a standard way of measuring the global importance or discriminative power of terms in text. When IDF is used in combination with term frequency (TF), the result is a very robust and highly effective term weighting scheme that has been applied across a

wide range of application areas, including databases, knowledge management, text classification, natural language processing, and, of course, information retrieval.

Sparck Jones' original proposal for IDF was based on empirical observations of global term frequency, whereby highly frequent terms should be given less weight than less frequent terms since they are more common and less discriminative. Two years after the original proposal, Robertson and Sparck Jones derived the so-called binary independence retrieval (BIR) model [16], which is more appropriately called the linked dependence model [3]. Later, Croft and Harper [4] showed that, under a number of assumptions, IDF can be derived directly from the BIR model in the form of a RSJ relevance weight. A similar line of research was also explored by Robertson and Walker [19], who investigated the relationship between RSJ relevance weights and IDF in more depth. Other derivations of IDF have been proposed [1, 13], but the one that is most appealing, especially from an information retrieval point of view, is the one based on the RSJ weighting. Therefore, it is this derivation that is the primary focus of our investigation in this paper.

In the field of information retrieval, there have been many attempts to change how the TF component in TF.IDF term weighting is computed [8, 18, 21]. However, there has been few, if any, attempts to improve upon the small number of "classical" IDF formulations. This may be the case because it is non-trivial to change the standard IDF formulation in a theoretically meaningful way while improving effectiveness. There may be heuristic ways to alter the IDF formulation, but doing so leads to little in the way of improved understanding as to *why* things improved. In this paper, we investigate, at a very fundamental level, the various assumptions that are entwined in the RSJ weighting derivation of IDF. By digging deeply into these assumptions, we are able to develop a better understanding of how and why IDF works. This analysis also allows us to propose a new, generalized form of IDF that we call *generalized IDF* (GIDF).

Our proposed generalized IDF results from taking a Bayesian view of the BIR model and decomposing the statistical assumptions. We not only examine the statistical assumptions made by previous researchers with respect to IDF, but also propose a new set of assumptions that lead to new IDF formulations. Our assumptions and formulations are backed up by empirical data that suggest the classical IDF formulations are non-optimal and can be improved upon.

This paper has four primary contributions. First, we provide a novel look at the BIR model in terms of a hierarchical Bayesian model. Second, we derive a generalized IDF for-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10 ...\$5.00.

mulation based on the model. Third, we propose a number of new IDF formulations based on GIDF that are inspired by data analysis and previous research, including an IDF that does not depend on document frequency. Finally, we test the effectiveness of our new IDF formulations on a number of TREC *ad hoc* retrieval data sets. Our results show that our proposed IDF formulations can consistently and significantly improve effectiveness over a number of classical IDF formulations.

The remainder of this paper is laid out as follows. First, in Section 2 we describe related work, and discuss several previously proposed derivations, including the classical derivation as a Robertson-Sparck Jones relevance weight. Then, in Section 3 we derive our generalized IDF, show how it relates to classical definitions of IDF, and propose several new IDF measures based on our derivation. In Section 4, we compare and contrast classical definitions of IDF with several instantiations of our proposed generalized IDF and show that our newly proposed IDF measures can consistently improve upon classical IDF measures on several *ad hoc* retrieval test collections. Finally, in Section 5 we conclude the paper and describe various potential areas of future work.

2. RELATED WORK

In this section we review the previous research that has laid the theoretical and empirical foundations for the current understanding of IDF. Later, we will build upon this foundation to develop a generalized formulation for IDF.

2.1 IDF from a Classical IR Perspective

We begin our discussion of related work by revisiting various theoretical derivations of IDF. We primarily focus on the derivation of IDF as a RSJ relevance weight and briefly describe other theoretical derivations that have been proposed.

As we discussed in the introduction, IDF can be theoretically derived as a relevance weight within Robertson and Sparck Jones’ BIR model [16]. Within the model, documents d are ranked according to $P(R = 1|D = d)$. Here, R is a binary random variable that represents relevance. Furthermore, D is a random variable that represents a document. Thus, the model ranks documents in decreasing likelihood of being relevant, which adheres to the well-known Probability Ranking Principle [14].

The model assumes that $d \in \{0, 1\}^{\mathcal{V}}$ is a random binary vector which has one entry for every term in the vocabulary \mathcal{V} . Here $d_i = 1$ if and only if term i occurs in the document. The standard BIR model ranking function derivation is as follows¹:

$$\begin{aligned} P(r|d) &= \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|r)P(r)}{P(d_i|\bar{r})P(\bar{r})} \\ &\stackrel{rank}{=} \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|r)}{P(d_i|\bar{r})} \\ &\stackrel{rank}{=} \sum_{i:d_i=1} \log \frac{P(d_i|r)P(\bar{d}_i|\bar{r})}{P(\bar{d}_i|r)P(d_i|\bar{r})} \end{aligned}$$

¹For notational convenience, we will use r and \bar{r} as shorthand for $R = 1$ and $R = 0$, respectively, and d as shorthand for $D = d$, as well.

where $wt_i = \log \frac{P(d_i|r)P(\bar{d}_i|\bar{r})}{P(\bar{d}_i|r)P(d_i|\bar{r})}$ is known as the Robertson-Sparck Jones (RSJ) relevance weight. This derivation invokes a number of assumptions. First, it is explicitly assumed that $P(d|r) = \prod_{i=1}^{|\mathcal{V}|} P(d_i|r)$, which imposes an independence assumption on the term occurrences. It is also implicitly assumed that $P(D|r)$ is distributed according to a multivariate Bernoulli distribution (i.e., each $P(d_i|r)$ is distributed as a Bernoulli). This assumption is implicit due to the fact that documents are represented as binary vectors. More recently, other distributional assumptions have been explored in contexts similar to the BIR model, including the Poisson [5, 7, 17, 20], multinomial [11], and Dirichlet compound multinomial [23] distributions, although most of the widespread IDF formulations are based on the multivariate Bernoulli distribution.

To use the BIR model, one must estimate $P(d_i|r)$ and $P(d_i|\bar{r})$ for every term i in the vocabulary for a total of $2|\mathcal{V}|$ parameters. It is feasible to estimate the probabilities if relevance information is available. For example, one may use the estimates originally proposed by Robertson and Sparck Jones, which are:

$$\begin{aligned} P(d_i|r) &= \frac{c(d_i, R)}{|R|} \\ P(d_i|\bar{r}) &= \frac{c(d_i, C) - c(d_i, R)}{|C| - |R|} \end{aligned}$$

where $c(d_i, R)$ is the number of relevant documents that term d_i occurs in, $c(d_i, C)$ is the number of documents that term d_i occurs in, $|R|$ is the total number of relevant documents, and $|C|$ is the total number of documents in the collection. This results in the following RSJ relevance weight:

$$wt_i = \log \frac{c(d_i, R) \cdot (|C| - |R| - c(d_i, C) + c(d_i, R))}{(|R| - c(d_i, R)) \cdot (c(d_i, C) - c(d_i, R))} \quad (1)$$

In many retrieval scenarios it is difficult or impossible to determine $|R|$ and $c(d_i, R)$ for each term. This makes the BIR model, as is, difficult to use in practice. Later, Croft and Harper proposed two assumptions that would allow the model to be used when no relevance information was available [4]. The first assumption they propose states that $P(d_i|\bar{r}) \approx P(d_i|C) = \frac{c(d_i, C)}{|C|}$, which says that the probability of any term occurring in the non-relevant class of documents is approximately equal to the probability of that term occurring in the entire collection. Their second assumption states that $P(d_i|r) = \pi$ for every term i that occurs in the query. Under these two assumptions, the RSJ relevance weight for term i reduces to:

$$wt_i = \log \frac{\pi}{1 - \pi} + \log \frac{N - df}{df} \quad (2)$$

It should also be noted that a similar IDF formulation can be derived from the RSJ relevance weight in Equation 1 if we assume that $c(d_i, R) = |R| = 0$. After adding the commonly used 0.5 “correction”, the following IDF formulation is achieved:

$$wt_i = \log \frac{|C| - c(d_i, C) + 0.5}{c(d_i, C) + 0.5} \quad (3)$$

which we will refer to as the *RSJ IDF* formulation.

The first assumption made by Croft and Harper is realistic and tends to hold in practice, as most of the documents in a reasonably sized collection will not be relevant to a given

query. The second assumption, however, is an oversimplification. In practice, as others have shown [6], and as we will also show, $P(d_i|r)$ tends to increase for very frequent terms. This makes intuitive sense, because very common terms are at least as likely as less common terms to occur in relevant documents. In addition, the IDF derived as the result of imposing the assumptions can result in *negative* IDF values for very frequent terms, which is undesirable, both from a theoretical and practical point of view.

These factors led Robertson and Walker to propose a new estimate for $P(d_i|r)$ that increases as the frequency of term i increased within the collection [19]. Their proposed estimate has the form:

$$P(d_i|r) = \frac{\pi}{\pi + (1 - \pi) \frac{|C| - c(d_i, C)}{|C|}}$$

where π is the same constant used in the Croft and Harper estimate. However, $P(d_i|r) = \pi$ only when $c(d_i, C) = 0$. Furthermore, $P(d_i|r) = 1$ if $c(d_i, C) = |C|$. Therefore, the estimate is more in line with empirical observations. Using this estimate and the 0.5 ‘‘correction’’, the RSJ relevance weight becomes:

$$wt_i = \log \frac{|C| + 0.5}{c(d_i, C) + 0.5} \quad (4)$$

which we will call the *RSJ positive IDF*. Indeed, not only does this estimate tend to model the actual data better, but it also tends to lead to better retrieval effectiveness than other IDF formulations derived as a RSJ relevance weight.

Recently, Lee proposed a new estimate for $P(d_i|r)$ that is similar to Laplacian smoothing [12]. The proposed estimate has the form:

$$P(d_i|r) = \frac{c(d_i, C) + L}{|C| + L}$$

where L is a smoothing parameter that controls how many pseudo-counts get added to both the document frequency as well as the number of documents. This estimate results in the following RSJ weight:

$$wt_i = \log \left(1 + \frac{L}{c(d_i, C)} \right) \quad (5)$$

It is easy to see that when $L = N$, the formulation takes on the standard IDF form. Lee suggests that replacing the constant L with a function $L(i)$ that depends on the term may be useful for constructing a more effective IDF.

As should now be clear, most of the work that has gone into improving IDF has looked at new estimates for $P(d_i|r)$. Even in this vein of research there has been very little done, and very few experimental results to show for it. Furthermore, the estimate for $P(d_i|\bar{r})$ has been largely ignored. Thus, the goal of our work is to consider various general estimates for both $P(d_i|r)$ and $P(d_i|\bar{r})$ and rigorously evaluate the effectiveness of each in order to develop an even better understanding of how the various components of IDF interact with each other.

2.2 Other Derivations of IDF

Although we have primarily focused on IDF, as derived from the BIR model, there have been other derivations proposed in the literature, almost exclusively from an information theoretic point of view. These include the derivation proposed by Aizawa [1] and Papineni [13], both of which

made theoretical arguments based on information theoretic principles. These derivations are appealing because they use different tools and machinery to derive the same notion that was originally derived using information retrieval ideas and concepts. Therefore, rather than viewing these derivations as competing, it is useful to glean as much from each in order to develop a deeper understanding of IDF from multiple perspectives.

3. GENERALIZED IDF

We now describe the details of our proposed generalized IDF formulation. Our derivation and subsequent estimates are similar in nature to those proposed previously. However, we take a fresh view of things and propose novel estimates that go beyond those that are found in the literature. Our proposed estimates are not only theoretically well-founded, but also highly effective, as we will show later. The general framework laid out in the remainder of this section can be used to easily derive new IDF variants that are more flexible and robust than the current variants and lead to improved retrieval effectiveness for a wide variety of tasks.

3.1 Derivation

The original BIR model does not explicitly have a random variable for modeling the query. Everything is assumed to be implicitly conditioned on some information need. However, it is often more convenient and more flexible to include the query within the model. Lafferty and Zhai showed that the BIR model can easily be generalized to include a query random variable Q [10]. We build upon their work to formally derive our generalized form of IDF as follows:

$$\begin{aligned} RSV(q, d, J) &= \frac{P(r, |d, q, J)}{P(\bar{r}, |d, q, J)} \\ &= \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|q, r, J)P(q, r, J)}{P(d_i|q, \bar{r}, J)P(q, \bar{r}, J)} \\ &\stackrel{rank}{=} \log \prod_{i=1}^{|\mathcal{V}|} \frac{P(d_i|q, r, J)}{P(d_i|q, \bar{r}, J)} \\ &\stackrel{rank}{=} \sum_{i:d_i=1} \log \frac{P(d_i|q, r, J)P(\bar{d}_i|q, \bar{r}, J)}{P(\bar{d}_i|q, r, J)P(d_i|q, \bar{r}, J)} \end{aligned}$$

where Q is a random variable that represents the query or information need and q is a specific query in the event space of Q , and J is set of relevance information. Here, J , which is novel to our derivation, stands for ‘judgments’, and can be any source of relevance data available, including (pseudo-)relevance feedback documents or click-through logs. The set J is partitioned into J_r and $J_{\bar{r}}$, which are the relevant and non-relevant judgment data in J , respectively. These sets provide evidence when computing the probability of relevance (or non-relevance) for a given query. From this derivation, we obtain the following RSJ-like relevance weight:

$$wt_i = \log \underbrace{\frac{P(d_i|q, r, J)}{P(\bar{d}_i|q, r, J)}}_{IDF_r} + \log \underbrace{\frac{P(\bar{d}_i|q, \bar{r}, J)}{P(d_i|q, \bar{r}, J)}}_{IDF_{\bar{r}}}$$

which we can split into a relevant component (IDF_r) and a non-relevant component ($IDF_{\bar{r}}$). As we will show, different

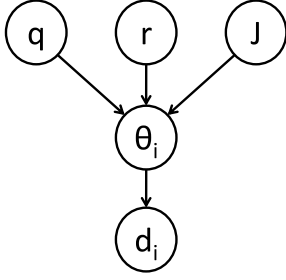


Figure 1: Graphical model representation of the generalized probabilistic model.

instantiations of our generalized IDF will result in different formulations for these two components.

Thus, we must estimate $P(d_i|q, r, J)$ and $P(d_i|q, \bar{r}, J)$ for every term i in the vocabulary. Although this is very similar in nature to the estimation problem encountered in the original BIR model, it is unique in the sense that we now have two additional pieces of evidence (q and J) from which the probabilities can be estimated more effectively. Unlike the BIR model, which models $P(d_i|r)$ and $P(d_i|\bar{r})$ according to multivariate Bernoulli distributions, we take a hierarchical Bayesian approach to estimate $P(d_i|q, r, J)$ and $P(d_i|q, \bar{r}, J)$. The graphical model representation of our hierarchical model is shown in Figure 1. Under the model, for each term i , a query q , relevance setting r , and a set of relevance information J imposes a probability distribution θ_i from which the term i is sampled for each document in the collection.

We assume that $P(d_i|\theta)$ is distributed according to a Bernoulli distribution and $P(\theta|q, r, J)$ is distributed according to a Beta distribution with hyperparameters $\alpha_i(q, r, J)$ and $\beta_i(q, r, J)$. One particularly nice property of the Beta distribution is that it is a conjugate prior to the Bernoulli. In addition, as we will show, it models the actual distribution of the model parameters well in practice. Thus, under these assumptions, our estimates can be computed as follows:

$$P(d_i|q, r, J_r) = \int_{\theta} P(d_i|\theta)P(\theta|q, r, J)d\theta$$

$$P(d_i|q, \bar{r}, J_{\bar{r}}) = \int_{\theta} P(d_i|\theta)P(\theta|q, \bar{r}, J)d\theta$$

Therefore, the estimates are simply expectations over θ , which can be efficiently computed as:

$$P(d_i|q, r, J) = \mathbb{E}[\theta|q, r, J]$$

$$= \frac{\alpha_i(q, r, J)}{\alpha_i(q, r, J) + \beta_i(q, r, J)}$$

$$P(d_i|q, \bar{r}, J) = \mathbb{E}[\theta|q, \bar{r}, J]$$

$$= \frac{\alpha_i(q, \bar{r}, J)}{\alpha_i(q, \bar{r}, J) + \beta_i(q, \bar{r}, J)}$$

since the mean of a Beta random variable with hyperparameters α and β is $\frac{\alpha}{\alpha+\beta}$.

Using these estimates, we obtain the following:

$$wt_i = \underbrace{\log \frac{\alpha_i(q, r, J)}{\beta_i(q, r, J)}}_{IDF_r} + \underbrace{\log \frac{\beta_i(q, \bar{r}, J)}{\alpha_i(q, \bar{r}, J)}}_{IDF_{\bar{r}}} \quad (6)$$

We call this formulation the *generalized IDF*, since it provides a great deal of flexibility and generalizes most of the previously proposed IDF formulations. With the original RSJ relevance weight-based IDF formulation, the problem boiled down to estimating $P(d_i|r)$ and $P(d_i|\bar{r})$. In this case, we are still estimating probabilities, but we have shifted our focus from directly estimating the probabilities, to finding reasonable settings of the model hyperparameters α and β . As we will now show, this is not a difficult task, and it is quite simple to come up with a range of new, easily interpretable IDF formulations based on the generalized IDF formula.

3.1.1 Relevance Information

If we have relevance information (i.e., $J \neq \{\}$) in the form of judged documents, then there are many reasonable ways of setting the α and β parameters. For example, one possible setting is the following:

$$\alpha_i(q, r, J) = c(d_i, J_r)$$

$$\beta_i(q, r, J) = |J_r| - c(d_i, J_r)$$

$$\alpha_i(q, \bar{r}, J) = c(d_i, C) - c(d_i, J_r)$$

$$\beta_i(q, \bar{r}, J) = |C| - |J_r| - c(d_i, C) + c(d_i, J_r)$$

where $|J_r|$ is the number of documents in J_r and $c(d_i, J_r)$ is the number of documents in J_r that term i occurs. These settings result in the following IDF estimate:

$$wt_i = \log \frac{c(d_i, J_r)}{|J_r| - c(d_i, J_r)} + \log \frac{|C| - |J_r| - c(d_i, C) + c(d_i, J_r)}{c(d_i, C) - c(d_i, J_r)} \quad (7)$$

which is the original RSJ relevance weight without the 0.5 “correction” (Equation 1). The 0.5 “correction” that is commonly used can be obtained by adding 0.5 to all of the hyperparameters above. This biases the mean of $P(\theta|q, r, J)$ away from $\frac{c(d_i, J_r)}{|J_r|}$, but typically results in improved retrieval effectiveness.

Thus, we have derived the RSJ relevance weight with relevance information using our generalized IDF framework. Indeed, it is easy to derive other relevance-based IDF formulations within this framework by simply changing the model hyperparameters. We hope this will stimulate further research into new and more robust IDF formulations, especially now that novel types of relevance information, such as click-through data, is available.

We do not run any experiments using relevance information in this work. Therefore, this derivation is more for theoretical completeness. As part of future work we plan to carry out experiments that validate the effectiveness and robustness of this estimate versus the original RSJ relevance weight.

3.1.2 No Relevance Information

The more interesting, and common, case arises when we have no relevance information available (i.e., $J = \{\}$). There are two possible cases to consider here. First, we may assume, in a similar fashion to Croft and Harper, that every document in the entire collection is not relevant. This is equivalent to setting $J_{\bar{r}} = C$. Although the Croft and Harper assumption has been shown to hold in practice, we

do not feel that it is theoretically well-founded, since we have not actually observed any non-relevant documents. We believe that the Croft and Harper assumption should be encoded directly into the model prior, since this is *a priori* knowledge, rather than actual observed evidence. Therefore, we argue that a more proper formulation simply treats $J_{\bar{r}}$ as empty and enforces any *a priori* knowledge into the prior over θ via the hyperparameters α and β . In the next two sections we describe various ways of setting the hyperparameters when no relevance information is available.

3.2 Relevance Distribution Assumptions

We begin by examining various ways of formulating IDF_r , which is the part of the IDF formula that is based on the relevant class of documents. Since we only consider generalized IDF, formulating IDF_r is equivalent to finding settings for $\alpha_i(q, r)$ and $\beta_i(q, r)$ which determine the distribution of models in the relevant class. Notice that we have dropped J from the hyperparameter formulation, since it is assumed to be empty.

We now describe two sets of assumptions over $\alpha_i(q, r)$ and $\beta_i(q, r)$, their corresponding IDF_r formulations, and their statistical interpretation within the model.

3.2.1 Assumption Set 1

Our first assumption, which is very simple, states that IDF_r is a constant. This is achieved by making the following assumptions for $\alpha_i(q, r)$ and $\beta_i(q, r)$:

$$\frac{\alpha_i(q, r)}{\beta_i(q, r)} = \frac{\gamma}{1 - \gamma}$$

where $\gamma \in (0, 1)$ is a free parameter. The smaller γ is, the smaller the ratio of $\alpha_i(q, r)$ to $\beta_i(q, r)$ is, which results in more pseudo-counts being assigned to the non-occurrences of the term than occurrences of the term in the prior. This results in the following IDF_r formulation:

$$IDF_r = \frac{\gamma}{1 - \gamma}$$

Perhaps a simpler, more straightforward interpretation of this assumption can be gleaned from the resulting estimate for $P(d_i|q, r)$, which simply is:

$$P(d_i|q, r) = \gamma$$

Therefore, this assumption states that $P(d_i|q, r)$ is *constant* for all terms. This is equivalent to Croft and Harper’s assumption, which is known to be inaccurate [6]. Indeed, for the data sets that we consider in this paper, this assumption also is inaccurate. Figure 2 plots the probability that a word occurs in a relevant document versus the probability that the word occurs in any document. Each point represents a single term in a query. The plots are generated across large set of queries. As the plot shows, there is a noticeable trend in the data that discredits the Croft and Harper assumption. Indeed, the plots support previous research that suggests $P(d_i|q, r)$ is increasing as a function of $P(d_i|C)$.

3.2.2 Assumption Set 2

Given this evidence, we seek to formulate an $P(d_i|q, r)$ that increases as $P(d_i|C)$ increases. There are many ways that this can be done within the generalized IDF framework. However, for the purpose of this paper, we consider the the

following assumption:

$$\begin{aligned} \alpha_i(q, r) &= (1 - \lambda) \cdot \mathbb{E}[\theta|r] + \lambda P(d_i|C) \\ \beta_i(q, r) &= 1 - \alpha_i(q, r) \end{aligned}$$

where $\mathbb{E}[\theta|r]$ is the average θ (i.e., $P(d_i|q, r)$) across all query terms in our training set. This represents the “average” probability of a term occurring in a relevant document. In addition, $P(d_i|C) = \frac{c(d_i, C)}{|C|}$, is the probability of the term occurring in the entire collection. This results in the following IDF_r formulation:

$$IDF_r = \log \frac{(1 - \lambda) \cdot \mathbb{E}[\theta|r] \cdot |C| + \lambda c(d_i, C)}{|C| - \lambda c(d_i, C) - (1 - \lambda) \cdot \mathbb{E}[\theta|r] \cdot |C|}$$

and the following probability estimate:

$$P(d_i|q, r) = (1 - \lambda) \cdot \mathbb{E}[\theta|r] + \lambda P(d_i|C)$$

Therefore, the estimate smooths the collection probability with the average probability that any term will occur in a relevant document. As long as $\lambda > 0$, this estimate will increase as $P(d_i|C)$ increases, thereby matching the empirical observations.

This estimate is similar to the popular Jelinek-Mercer smoothing from statistical language modeling. There are many other possible ways to interpolate, or smooth, the collection probability against the sample mean, including Laplacian smoothing, absolute discounting, and Dirichlet smoothing [24].

3.3 Non-relevance Distribution Assumptions

We now investigate four assumptions for $\alpha_i(q, \bar{r})$ and $\beta_i(q, \bar{r})$. As with the relevance distribution assumptions, by setting these parameters, we are imposing a form on $IDF_{\bar{r}}$, which is the part of the IDF formula that deals with the distribution of non-relevant documents.

As before, we describe each assumption, the corresponding $IDF_{\bar{r}}$ formulations, and the statistical properties surrounding the assumption.

3.3.1 Assumption Set 1

Our first assumption for the non-relevant distribution is exactly the same as our assumption for the relevant distribution, and is as follows:

$$\frac{\alpha_i(q, \bar{r})}{\beta_i(q, \bar{r})} = \frac{\gamma}{1 - \gamma}$$

which results in an analogous formulation for $IDF_{\bar{r}}$:

$$IDF_{\bar{r}} = \frac{\gamma}{1 - \gamma}$$

as well as an analogous probability estimate:

$$P(d_i|q, \bar{r}) = \gamma$$

which, as before, is constant for every term in the vocabulary. Of course, this is a poor assumption, but the formulation does not depend on any collection statistics, such as document frequency. Indeed, when this assumption is used in conjunction with assumption set 1 from the relevant class, we obtain a formulation of IDF that does not depend on the document frequency in any way. Since the same weight is assigned to every query term, this is equivalent to using no IDF information whatsoever. Therefore, we expect this particular combination of assumption sets to perform poorly.

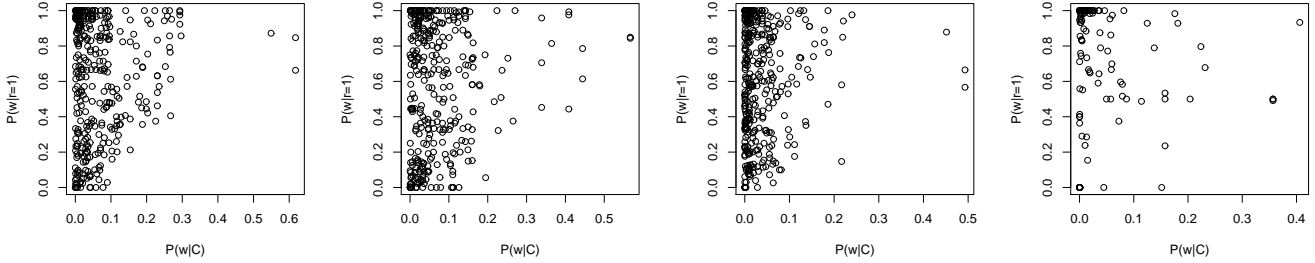


Figure 2: Probability that a word occurs in a judged relevant document ($P(d_i|q, r)$) plotted as a function of the probability the word occurs in *any* document ($P(d_i|C)$) across four TREC data sets.

3.3.2 Assumption Set 2

Our first assumption set for the non-relevant class of documents attempts to mimic the Croft and Harper $IDF_{\bar{r}}$, by using the following settings for $\alpha_i(q, \bar{r})$ and $\beta_i(q, \bar{r})$:

$$\begin{aligned}\alpha_i(q, \bar{r}) &= c(d_i, C) + \gamma \\ \beta_i(q, \bar{r}) &= |C| - c(d_i, C) + \gamma\end{aligned}$$

where, as before, γ is a free parameter. We must set γ as to ensure $\alpha_i(q, \bar{r})$ and $\beta_i(q, \bar{r})$ remain positive to ensure the Beta distribution remains well-defined. In our experiments, we only consider $\gamma \geq 0$. This setting of the hyperparameters results in the following $IDF_{\bar{r}}$:

$$IDF_{\bar{r}} = \log \frac{|C| - c(d_i, C) + \gamma}{c(d_i, C) + \gamma}$$

which, as we had hoped for, is reminiscent of Croft and Harper’s formulation. Indeed, if we set $\gamma = 0.5$, then we obtain the commonly used 0.5 “correction”. Here, it is very easy to see exactly where the 0.5 comes from and what its meaning actually is. The actual probability estimate for $P(d_i|q, \bar{r})$ using this formulation is:

$$P(d_i|q, \bar{r}) = \frac{c(d_i, C) + \gamma}{|C| + 2\gamma}$$

This estimate degenerates to $P(d_i|C)$ when $\gamma = 0$ and converges towards $\frac{1}{2}$ as $\gamma \rightarrow \infty$. However, this estimate still suffers from the problem of resulting in negative $IDF_{\bar{r}}$ values for highly frequent terms, which is undesirable and can hurt retrieval effectiveness.

3.3.3 Assumption Set 3

Therefore, to overcome the problem of problem of negative $IDF_{\bar{r}}$ values, we propose the following settings for the Beta hyperparameters:

$$\begin{aligned}\alpha_i(q, \bar{r}) &= c(d_i, C) + \gamma \\ \beta_i(q, \bar{r}) &= |C| + \gamma\end{aligned}$$

Notice that the only difference between this assumption set and the previous one is that $c(d_i, C)$ is no longer subtracted from $\beta_i(q, \bar{r})$, thereby eliminating the possibility for $\beta_i(q, \bar{r}) > \alpha_i(q, \bar{r})$, which is the cause for the negative values. This slightly modified assumption set results in the following $IDF_{\bar{r}}$ formulation:

$$IDF_{\bar{r}} = \log \frac{|C| + \gamma}{c(d_i, C) + \gamma}$$

and the following probability estimate:

$$P(d_i|q, \bar{r}) = \frac{c(d_i, C) + \gamma}{|C| + c(d_i, C) + 2\gamma}$$

3.3.4 Assumption Set 4

Our final assumption set for the non-relevant distribution mirrors assumption set 2 from the relevance distribution. The assumptions is as follows:

$$\begin{aligned}\alpha_i(q, \bar{r}) &= ((1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] + \lambda P(d_i|C)) \\ \beta_i(q, \bar{r}) &= 1 - \alpha_i(q, \bar{r})\end{aligned}$$

where $\mathbb{E}[\theta|\bar{r}]$ is the average θ (i.e., $P(d_i|q, \bar{r})$) across all terms and queries in our training set. Rather than estimating this from the set of judged non-relevant documents, which is very biased, we estimate this from the entire collection. The resulting $IDF_{\bar{r}}$ formulation is then:

$$IDF_{\bar{r}} = \log \frac{|C| - \lambda c(d_i, C) - (1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] \cdot |C|}{(1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] \cdot |C| + \lambda c(d_i, C)}$$

and our probability estimate is:

$$P(d_i|q, \bar{r}) = (1 - \lambda) \cdot \mathbb{E}[\theta|\bar{r}] + \lambda P(d_i|C)$$

which, as before, is a linear combination of the collection probability and the sample mean observed in the training data. Different settings of λ result in different estimates and control the amount by which $P(d_i|q, \bar{r})$ increases with $P(d_i|C)$.

3.4 Relationship to Classical Formulations

It is easy to derive the classical IDF formulations within the generalized IDF framework using the various assumption sets proposed here. By interpreting classical IDFs in this way, we are able to shed new insights into how these formulations are related to each other and also motivate new and interesting directions for further research into improved IDF measures.

The RSJ IDF, as defined above, can be derived using generalized IDF using assumption set 1 for IDF_r with $\gamma = 0.5$ and assumption set 2 $IDF_{\bar{r}}$ with $\gamma = 0.5$. Furthermore, the RSJ positive IDF, as defined above, can also be derived using assumption set 1 for IDF_r with $\gamma = 0.5$ and assumption set 3 $IDF_{\bar{r}}$ with $\gamma = 0.5$. Although the formulation proposed by Lee cannot be derived using any of our proposed assumption sets, it is easy to see that it is very similar to using assumption set 2 for IDF_r and assumption set 2 for

Name	Description	# Docs	Train Topics	Test Topics
WSJ	Wall St. Journal 87-92	173,252	51–150	151–200
AP	Assoc. Press 88-90	242,918	51–150	151–200
ROBUST	Robust 2004 data	528,155	301–450	601–700
WT10G	TREC Web collection	1,692,096	451–500	501–550

Table 1: Overview of TREC collections and topics.

$IDF_{\bar{r}}$. It is easy to use the generalized IDF framework to derive the exact IDF formulation by introducing a new assumption set for IDF_r .

Therefore, our generalized IDF is truly general, in the sense that most of the previously proposed IDF measures that have been derived as a RSJ relevance weight, can easily be derived within the framework.

4. EMPIRICAL EVALUATION

We now describe our empirical evaluation of our generalized IDF framework.

4.1 Data and Experimental Setup

Our experiments are conducted over four standard TREC *ad hoc* retrieval data sets. The data sets are summarized in Table 1. The WSJ, AP, and ROBUST data sets are medium size and consist solely of news articles, whereas the WT10G data set is much larger and contains web documents.

Each data set is divided into a training and a test set. For all of our experiments, we only use the training set for tuning model parameters. However, we generally report both training and test set effectiveness for the sake of completeness.

Our evaluation methodology follows the standard TREC procedures. Our primary evaluation metric of interest is mean average precision computed to a depth of 1000 documents per query. All training is done to directly maximize mean average precision. All statistical significance tests reported use of a one-tailed *t*-test.

All of the experiments were carried out using the soon to be released Searching using Markov Random Fields (SMRF) toolkit, which provides a robust experimental layer that sits on top of an Indri index [22]. All documents are stopped using a standard stopword list and stemmed using the Porter stemmer.

4.2 Empirical Analysis

Before describing our *ad hoc* retrieval experiments, we first empirically analyze various characteristics of the TREC data sets.

First, we wish to determine how realistic it is to assume that $P(\theta|q, r)$ and $P(\theta|q, \bar{r})$ is distributed according to a Beta distribution. We chose this distribution because it is conjugate to the Bernoulli, which makes the mathematics and analysis simpler. However, this does not ensure that the Beta models the actual distributions well at all. Rather than proving that the fit is good using statistical analysis, we take a much simpler, higher level approach. We first plot a histogram of the observed $P(\theta|q, r)$ and $P(\theta|q, \bar{r})$ in

the data. Then, we fit a Beta distribution to the data. We can then see how good or bad the corresponding fit is.

The resulting histograms and fitted Beta distributions are plotted in Figure 3. The Beta distribution is fit to the empirical distribution by using the method of means, which sets the mean and variance of the Beta equal to the sample mean and variance. Given the mean and variance of the sample, the α and β hyperparameters are set as follows:

$$\alpha = \mathbb{E}[\theta] \cdot \left(\frac{\mathbb{E}[\theta] \cdot (1 - \mathbb{E}[\theta])}{\text{Var}[\theta]} - 1 \right)$$

$$\beta = (1 - \mathbb{E}[\theta]) \cdot \left(\frac{\mathbb{E}[\theta] \cdot (1 - \mathbb{E}[\theta])}{\text{Var}[\theta]} - 1 \right)$$

It is easy to verify that this setting yields a Beta distribution where the mean and variance are equal to the sample mean and variance, respectively. For illustrative purposes, the estimated α and β values for each of our data sets is provided in Table 2. It is interesting to note that $\mathbb{E}[\theta|r]$ tends to increase as the collection gets larger, and is especially large for the WT10G collection. This suggests that almost all of the relevant documents contain all of the query terms. A similar observation was also made by Buckley et al. [2]. For data sets with large $\mathbb{E}[\theta|r]$ we expect IDF_r to be large for most terms in order to promote a coordinate-level matching type of ranking function.

The fitted distributions in Figure 3 appear to be good fits to the empirical distributions. Therefore, the Beta assumption seems reasonable. However, it is important to note that this distribution is estimated over *all* terms, rather than individual terms, as is actually done with generalized IDF. However, since our data is so sparse, there is no reasonable way to do a similar analysis for each term. It would be interesting to revisit this analysis for a much larger training set in order to determine if each term actually follows a Beta distribution or not.

4.3 IDF Results

Our first set of experiments investigate how our generalized IDF formulations compare to each other and to classical IDF formulations using an entirely IDF-based ranking function of the form:

$$S(Q, D) = \sum_{i:d_i=1, q_i=1} (IDF_r + IDF_{\bar{r}})$$

It is well understood that IDF-only ranking functions are far inferior to TF.IDF ranking functions. However, by eliminating TF from the equation, we can focus on the effectiveness of the various IDF formulations in isolation. We will evaluate the formulations using a TF.IDF ranking function later, however.

Table 3 lists the results of our IDF-only experiments. For each data set, we provide both the training and test set mean average precision values for every combination of IDF_r and $IDF_{\bar{r}}$ assumption sets. We also include the classical RSJ and RSJ positive IDF formulations, which we showed earlier to be special cases of our generalized IDF framework. It is not our expectation to significantly improve upon the classical IDF formulations. Instead, the goal of these experiments is to show that the generalized IDF framework is very simple to use to construct new, potentially effective, IDF formulations.

In the table, bold values represent the best result within a given column. The first thing to notice is that one of the

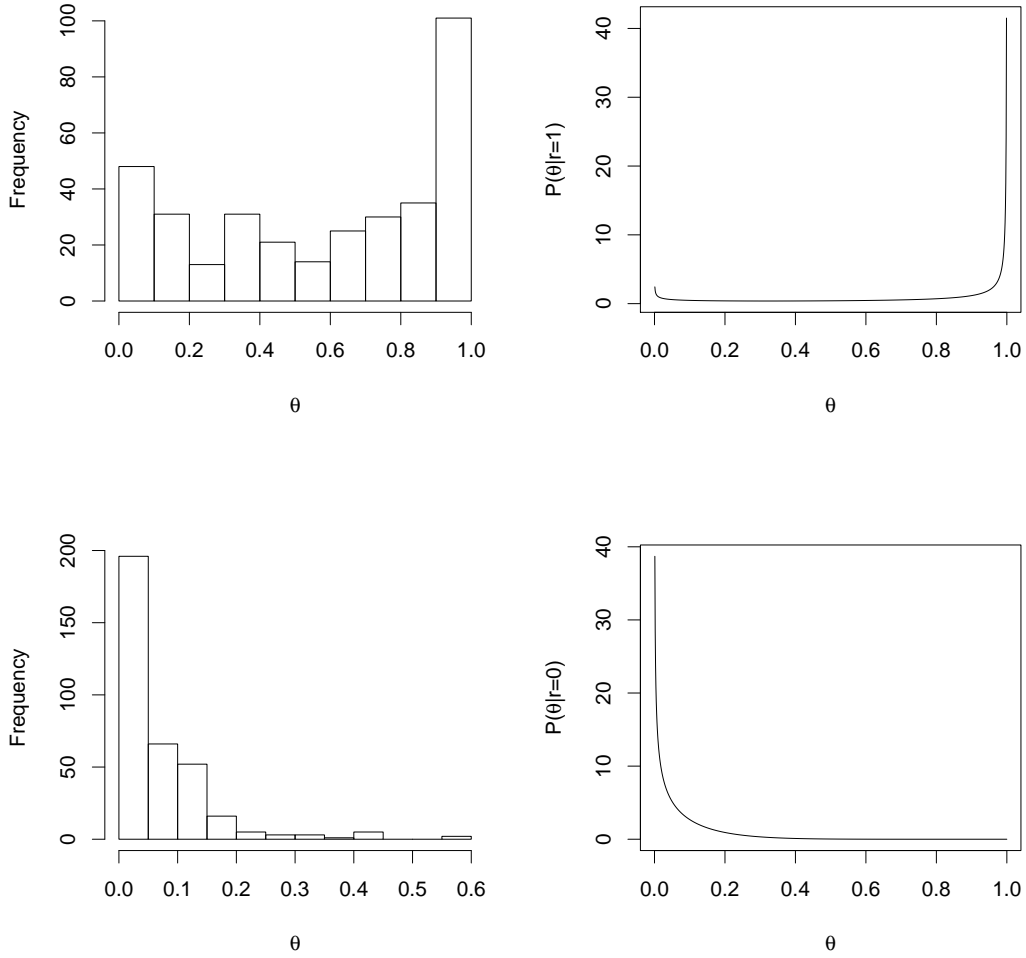


Figure 3: The figures in the first column show a histogram of the empirical distribution of θ observed in the training data for the relevant (top) and non-relevant (bottom) document classes in the AP data set. The figures in the second column show the Beta distribution that is fit to the histogram using the method of moments.

		Relevant				Non-Relevant			
		$\mathbb{E}[\theta r]$	$Var[\theta r]$	α	β	$\mathbb{E}[\theta \bar{r}]$	$Var[\theta \bar{r}]$	α	β
WSJ	Train	.6103	.1086	.7260	.4637	.0688	.0075	.5179	7.007
	Test	.5930	.1139	.6640	.4558	.0605	.0053	.5900	9.161
AP	Train	.5881	.1236	.5645	.3954	.0702	.0072	.5687	7.529
	Test	.5663	.1206	.5874	.4499	.0587	.0033	.9292	14.889
ROBUST	Train	.6108	.1001	.8397	.5352	.0366	.0036	.3262	8.575
	Test	.7230	.0952	.7974	.3056	.0370	.0022	.5647	14.69
WT10G	Train	.7434	.1035	.6270	.2165	.0429	.0062	.2429	5.422
	Test	.7380	.0986	.7089	.2517	.0476	.0043	.4589	9.182
ALL		.6321	.1124	.6756	.3932	.0539	.0052	.4743	8.327

Table 2: For each data set, as well as all of the data sets combined (ALL), we report the mean and variance of θ , as well as the corresponding α and β values using the method of moments for both the relevant and non-relevant classes.

Assumption		AP		WSJ		ROBUST		WT10G	
IDF_r	$IDF_{\bar{r}}$	Train	Test	Train	Test	Train	Test	Train	Test
1	1	.0601	.0758	.1000	.1328	.1123	.0996	.0494	.0367
1	2	.0767	.0096	.1312	.1782	.1416	.1257	.0621	.0555 ^{$\alpha\beta$}
1	3	.0771	.0969 ^{α}	.1308	.1779 ^{α}	.1417	.1257	.0625	.0594 ^{β}
1	4	.0765	.0963	.1305	.1782	.1416	.1257	.0620	.0553
2	1	.0601	.0758	.1000	.1328	.1123	.0996	.0494	.0367
2	2	.0775	.0971	.1314	.1830 ^{$\alpha\beta$}	.1421	.1254	.0623	.0556
2	3	.0778	.0969 ^{α}	.1313	.1829 ^{$\alpha\beta$}	.1421	.1254	.0623	.0556 ^{$\alpha\beta$}
2	4	.0765	.0971	.1307	.1837 ^{$\alpha\beta$}	.1421	.1254	.0621	.0555
RSJ		.0766	.0967	.1301	.1767	.1405	.1255	.0610	.0553
RSJ Positive		.0766	.0967	.1296	.1783	.1405	.1255	.0610	.0553

Table 3: IDF-only ranking function results. Training and test set mean average precision is reported for each combination of generalized IDF assumptions. Bold values indicate the best formulation for the each data set. The superscripts α and β indicate statistically significant improvements over RSJ and RSJ Positive, respectively, at the $p < 0.1$ level. Underlined superscripts are significant at the $p < 0.05$ level. Significance tests were only performed on the test sets.

generalized IDF formulations is always the best, both for the training and test sets. This result shows the robustness of the framework. Furthermore, the superscripts indicate statistically significant improvements over the classical IDF formulations. Such tests were only computed on the test sets, since that is the most meaningful analysis. We see that the generalized IDF is significantly better for a number of data sets, with relative improvements in mean average precision ranging up to 7%.

Unfortunately, no one formulation stands out as being the most effective across all data sets. However, the IDF_r assumption set 2 does appear to be consistently better than the IDF_r assumption set 1, which supports previous observations, as well [19].

4.4 TF.IDF Results

Now that we have shown that our proposed framework is effective when an IDF-only ranking function is used, we investigate whether or not using the new IDF formulations will be as effective when combined with TF in a TF.IDF ranking function. It is not immediately clear how, if at all, the TF component will interact with the IDF component.

For TF, we use the state of the art Okapi TF [18], which artfully normalizes for document length normalization and accounts for term frequency saturation. The Okapi TF is the TF component used in the widely used and highly effective BM25 ranking function [15]. The form of the Okapi TF that we use has the following form:

$$\hat{t}f(d_i, D) = \frac{(k_1 + 1) \cdot tf(d_i, D)}{k_1 \cdot (1 - b + b \cdot \frac{|D|}{|D|_{avg}}) + tf(d_i, D)}$$

where $tf(d_i, D)$ is the number of times that term d_i occurs in D , $|D|$ is the number of terms D , $|D|_{avg}$ is the average document length, and k_1 and b are free parameters, which are tuned on the training set.

We then combine the Okapi TF with our generalized IDF in the following way:

$$S(Q, D) = \sum_{i:d_i=1, q_i=1} \hat{t}f(d_i, D) \cdot (IDF_r + IDF_{\bar{r}})$$

The results of our TF.IDF experiments are shown in Table 4. As before, the bolded results indicate the best result

for the given column. As with the IDF-only ranking function, one of the generalized IDF formulations is always the most effective.

The results show that the TF.IDF model washes away some of the improvements in effectiveness that was observed with the IDF-only ranking function. However, the same general trends, observations, and analysis still holds. There still remain a number of results that are statistically significantly better than the classical IDF formulations, which suggests that the generalized IDF formulations proposed here can be used effectively in combination with powerful TFs, such as the Okapi TF.

Based on these results, and the results from the IDF-only ranking function, we would recommend using IDF_r assumption set 2 with $IDF_{\bar{r}}$ assumption set 4 as a good “out of the box” IDF. Of course, these IDF formulations require training unlike most of the classical formulations. However, there are two sides to the coin. While training data can be difficult to come by, it does exist for many tasks. If the data exists, then there is no reason why an uninformed IDF should be used, when a more general, robust IDF can be used instead.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we derived a new IDF formulation using a hierarchical Bayesian model that generalizes the classical BIR model. Our new IDF formulation, which can be interpreted as a RSJ relevance weight, is called generalized IDF. As we showed, it can be used to derive all of the classical IDF formulations and is extensible, allowing researchers to easily develop new IDF formulations based on empirical and theoretical analysis.

We evaluated the usefulness of our proposed framework by experimenting with eight derived IDF formulations. Experiments were carried out using both an IDF-only ranking function, as well as a TF.IDF ranking function. Our results showed consistent and significant improvements over classical IDF formulations on using both IDF-only and TF.IDF-based ranking functions. Our results also show that our proposed IDF is valuable and highly effective in isolation, but when intertwined with TF, some of the improvements are washed away, due to complex interactions with the TF component. This suggests that we extend our work to other

Assumption		AP		WSJ		ROBUST		WT10G	
IDF_r	$IDF_{\bar{r}}$	Train	Test	Train	Test	Train	Test	Train	Test
1	1	.1487	.1941	.2123	.2793	.1978	.2599	.1984	.1469
1	2	.1778	.2148	.2550	.3332	.2285	.2886	.2225	.1965 ^{β}
1	3	.1788	.2151	.2559	.3347 ^{$\alpha\beta$}	.2285	.2886	.2230	.1973 ^{β}
1	4	.1752	.2149	.2532	.3332	.2285	.2886	.2224	.1965 ^{β}
2	1	.1487	.1941	.2123	.2793	.1978	.2599	.1984	.1469
2	2	.1791	.2147	.2563	.3369^{$\alpha\beta$}	.2283	.2886	.2263	.2006^{$\alpha\beta$}
2	3	.1797	.2137	.2576	.3341	.2291	.2896	.2253	.1987 ^{β}
2	4	.1793	.2125	.2557	.3316	.2282	.2900	.2247	.2005 ^{$\alpha\beta$}
RSJ		.1778	.2149	.2550	.3332	.2279	.2892	.2225	.1965
RSJ Positive		.1781	.2147	.2147	.3332	.2283	.2888	.2217	.1947

Table 4: TF.IDF ranking function results. Training and test set mean average precision is reported for each combination of generalized IDF assumptions. Bold values indicate the best formulation for the each data set. The superscripts α and β indicate statistically significant improvements over RSJ and RSJ Positive, respectively, at the $p < 0.1$ level. Underlined superscripts are significant at the $p < 0.05$ level. Significance tests were only performed on the test sets.

distributions, such as the multinomial or Poisson, so that our derivations can naturally account for term frequency. We believe that such a model should be able to consistently and significantly improve over standard TF.IDF formulations, if modeled properly.

Furthermore, we plan to explore how the framework can be used for relevance and pseudo-relevance feedback. The framework has a natural mechanism for including both types of information. It would be interesting to compare the effectiveness of this new formulation against classical RSJ relevance weights estimated using relevance information.

6. REFERENCES

- [1] A. Aizawa. An information-theoretic perspective of TF-IDF measures. *Information Processing and Management*, 39(1):45–65, 2003.
- [2] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Inf. Retr.*, 10(6):491–508, 2007.
- [3] W. S. Cooper. Some inconsistencies and misnomers in probabilistic information retrieval. In *Proc. 14th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 57–61, New York, NY, USA, 1991. ACM.
- [4] W. B. Croft and D. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.
- [5] A. P. de Vries and T. Roelleke. Relevance information: a loss of entropy but a gain for IDF? In *Proc 28th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 282–289, New York, NY, USA, 2005. ACM.
- [6] W. R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proc. 21st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 11–19, New York, NY, USA, 1998. ACM.
- [7] S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26:197–206 and 280–289, 1975.
- [8] B. He and I. Ounis. On setting the hyper-parameters of term frequency normalization for information retrieval. *ACM Trans. Inf. Syst.*, 25(3):13, 2007.
- [9] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.
- [10] J. Lafferty and C. Zhai. Probabilistic relevance models based on document and query generation. In W. B. Croft and J. Lafferty, editors, *Language Modeling and Information Retrieval*. 2003.
- [11] V. Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts Amherst, Amherst, MA, 2006.
- [12] L. Lee. IDF revisited: a simple new derivation within the Robertson-Spärck Jones probabilistic model. In *Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 751–752, New York, NY, USA, 2007. ACM.
- [13] K. Papineni. Why inverse document frequency? In *Proc 2nd Proc. North American Chapter of the Assn. for Computational Linguistics on Language Technologies*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- [14] S. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.
- [15] S. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. 3rd Text REtrieval Conference*, pages 109–126, 1994.
- [16] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.
- [17] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proc. 3rd Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 35–56, Kent, UK, UK, 1981. Butterworth & Co.
- [18] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. 17th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
- [19] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proc. 20th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 16–24, New York, NY, USA, 1997. ACM.
- [20] T. Roelleke. A frequency-based and a poisson-based definition of the probability of being informative. In *Proc 26th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 227–234, New York, NY, USA, 2003. ACM.
- [21] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. 19th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.
- [22] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.
- [23] Z. Xu and R. Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *Proc. 31st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page To appear., 2008.
- [24] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.