The D2Q2 Framework: On the Relationship and Combination of Language Modelling and TF-IDF

Thomas Roelleke, Hany Azzam, Marco Bonzanini, Miguel Martinez-Alvarez and Mounia Lalmas

Abstract

Language Modelling (LM) and TF-IDF are two retrieval models with different foundations. There have been efforts aiming at establishing the relationship between these models, and whether one includes the other. Whether their combination could yield a third and better model is an open research question. This paper revisits the foundations of LM and TF-IDF and explores how these models' *bare* structures relate and how these structures can be combined. We begin with the premise that TF-IDF is the $P(d|q)/P(d)$ side of retrieval, which complements the common view that LM is $P(q|d)/P(q)$. Next, a hybrid framework based on the decomposition of the product of the two sides, $P(d|q)/P(d) \cdot P(q|d)/P(q)$, is developed. This leads to the *D2Q2* family of models, which joins the inner components of LM and TF-IDF instead of combining their scores. This paper provides new insights into the relationship between LM and TF-IDF, and experimental results show that the D2Q2 models perform comparably to competitive baselines.

1 Introduction

There has been significant research into how to combine retrieval models and how to relate them. Approaches such as [Bartell *et al.*, 1994; Croft *et al.*, 1990; Lee,] have shown the importance of combining different retrieval models through, for example, score fusion. Other approaches have proposed how to *analyse* different retrieval models' components and compare them [Fang and Zhai, 2005]. Both research directions have furthered the development of more effective models.

Two types of retrieval models that have been closely analysed and compared are language modelling (LM), and those based on term frequency (TF) and inverse document frequency (IDF). These models have different foundations. Variants of the former are based on the mixtures (smoothing) [Zhai and Lafferty, 2004; Zaragoza *et al.*, 2003]. TF-IDF models differ with regard to the TF quantification and normalisations employed [Robertson *et al.*, 1994; Singhal *et al.*, 1996; He and Ounis, 2005; Kwok, 1996; Taylor *et al.*, 2006]. Efforts to establish the relationship between these models and whether or not the former includes the features of the latter include [Zhai and Lafferty, 2001]. By examining the foundations of these retrieval models we learn that LM directly derives from the conditional probability $P(q|d)$ (q is the query, d is the document)

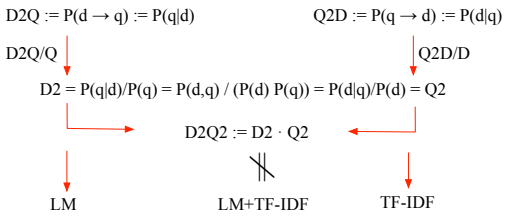

Figure 1: The D2Q2 framework.

[Ponte and Croft, 1998; Hiemstra, 2000; Lafferty and Zhai, 2003]. TF-IDF, on the other hand, is viewed as a *heuristic* model [Salton *et al.*, 1976; Croft, 2000; Metzler and Croft, 2004], and its probabilistic and information-theoretic interpretation is an ongoing debate [Church and Gale, 1995a; Church and Gale, 1995b; Aizawa, 2003; Robertson, 2004; Wu *et al.*, 2008; Roelleke and Wang, 2008]. Drawing from and furthering this type of deeper analysis allows us to better understand and relate these models' components.

This paper contributes several theoretical findings. We showcase a side-by-side derivation of LM and TF-IDF that helps to clarify the relationship between LM and TF-IDF. This derivation goes so far as to show that, just as LM has a TF-IDF nature, before the decomposition of document and query probabilities, *TF-IDF has an LM nature* as well. Next ,we develop a hybrid framework, leading to the D2Q2 family of models, that joins the inner components of LM and TF-IDF.

Figure 1 outlines the connections between D2Q2, its two subcomponents D2 and Q2, LM and TF-IDF. Essentially, D2Q2 rests on two ways to decompose the document-query independence measure DQI $:= P(d,q)/(P(d) \cdot P(q))$. D2Q2 combines the *inner* parts of TF-IDF and LM, trying to push the best of each into a *hybrid model*. Continuing with the derivation, the inner components of LM ($P(q|d)/P(q) = D2$) and TF-IDF ($P(d|q)/P(d) = Q2$) are combined. This provides an integrative framework that incorporates the characteristics of both LM and TF-IDF. Moreover, the instances of this framework, D2Q2, are retrieval models in their own right, which can be compared with the traditional models LM and TF-IDF. Interestingly, although D2Q2 "combines" models, it is different from the aforementioned fusion approaches. While fusion combines scores, D2Q2 incorporates the characteristics of both LM and TF-IDF into one probabilistic framework, and therefore we refer to D2Q2 as a "hybrid" model, as opposed to a model that fuses scores.

This paper is structured as follows. Section 2 consolidates the preliminaries necessary to appreciate the contri-

bution of this paper. Section 3 shows the relationship between LM and D2Q2. More precisely, it shows that LM corresponds to D2. The relationship between TF-IDF and D2Q2 is shown in Section 4 (TF-IDF corresponds to Q2). Section 5 discusses the relationship between LM and TF-IDF. From these follows in Section 6 the description of the D2Q2 framework, a theoretically sound combination of the LM and TF-IDF models into a family of hybrid models. Section 7 shows that the D2Q2 retrieval models perform comparably to competitive baselines.

2 Background & Preliminaries

2.1 LM and TF-IDF

We present the LM and TF-IDF models[1]. Note that TF-IDF is also referred to as a weighting scheme in the context of the vector space model. This paper emphasises that TF-IDF *is* a retrieval model at the same level as LM, as in [Hiemstra, 2000].

Let d be a document, q a query, c a collection and t a term. The standard definition of the retrieval status value associated with the LM model can be written as follows:

$$RSV_{\text{LM}}(d,q,c) := \sum_{t \in q} TF(t,q) \cdot \log\left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)}\right) \quad (1)$$

$TF(t,q)$ is the within-query term frequency, $P(t|d)$ is the within-document (foreground) term probability, $P(t|c)$ is the collection-wide (background) term probability, and λ_d is the document-dependent mixture parameter.

In the Dirichlet-based LM [Zhai and Lafferty, 2004], λ_d is proportional to the document length. Let $\lambda_d := \frac{dl}{dl+\mu}$, where *dl* is the document length and μ is a parameter. This setting of λ_d reflects trust in probabilities estimated from *long* documents.

$$RSV_{\text{Dirich-LM}}(d,q,c) := \sum_{t \in q} TF(t,q) \cdot \log\left(\frac{\mu}{\mu+dl} + \frac{dl}{dl+\mu} \cdot \frac{P(t|d)}{P(t|c)}\right) \quad (2)$$

The retrieval status value associated with the TF-IDF model can be written as follows:

$$RSV_{\text{TF-IDF}}(d,q,c) := \sum_{t \in d \cap q} TF(t,d) \cdot TF(t,q) \cdot IDF(t,c) \quad (3)$$

$TF(t,d)$ is the within-document term frequency quantification; $TF(t,q)$ is for the query. For independence of term occurrences, the setting is $TF(t,d) := tf_d$ where tf_d is the *total* within-document term frequency. This setting is known to be inferior to $TF(t,d) := tf_d/(tf_d + K_d)$, the setting known from BM25 [Robertson *et al.*, 1994], where K_d is a normalisation factor proportional to the pivoted document length, $pivdl(c) := dl/avgdl(c)$. We refer to this TF quantification as BM25-TF, and we also denote it as $TF_K(t,d)$, to make explicit the parameter K. For *IDF*, the common setting is $IDF(t,c) := -\log P_D(t|c)$, where $P_D(t|c) = df(t,c)/N_D(c)$ is the Document-based term probability (based on the set of Documents, hence, the subscript capital D), and $df(t,c)$ is the collection-wide document frequency of term t.

[1] Similar investigation was carried out for the BM25 model; however in this paper we focus on LM and TF-IDF.

Note that IDF is based on a *Document-based* term probability ($P(t|c) := P_D(t|c)$), whereas LM is *Location-based* ($P(t|c) := P_L(t|c)$) [Hiemstra, 2000]. We return to these two event spaces (Documents vs. Locations) in Section 4.5, where an essential assumption is made to establish the connection between TF-IDF and D2Q2.

2.2 Document-Query (In)dependence (DQI)

An common measure in probabilistic models is the document-query independence, formalised as follows:

$$DQI(d,q) := \frac{P(d,q)}{P(d) \cdot P(q)} \quad (4)$$

The DQI measures the document-query (in)dependence. $DQI = 1$ means that document and query intersect as if they were independent; $DQI < 1$ means that the intersection is less; and $DQI > 1$ means that the intersect is greater than if they were independent.

The *DQI* is a concept related to information theory. It is the inner component of the "mutual information" $MI(X,Y) := \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x) \cdot P(y)}$. The DQI is the argument of the log. The relationship of DQI to MI (and hence to conditional entropy) backs DQI as an information-theoretic measure [Gale and Church, 1991]. It also shows the theoretical justification of D2Q2, which leverages the DQI measure in its derivation. Lastly, DQI is related to exhaustiveness and specificity (another foundation of D2Q2).

2.3 Exhaustiveness and Specificity

The product $P(q|d) \cdot P(d|q)$ can be interpreted as *exhaustiveness* $\cdot$ *specificity*, where $P(q|d)$ is set to measure exhaustiveness and $P(d|q)$ specificity. These concepts were used in logic-based retrieval *frameworks* [Nie, 1992; Wong and Yao, 1995]. We retain the idea, and define an exhaustiveness-specificity measure:

$$ES(d,q) := P(q|d) \cdot P(d|q) \quad (5)$$

From this definition, it immediately follows the relationship between ES and DQI, which can be expressed in as follows:

$$ES(d,q) = \frac{P(d,q) \cdot P(d,q)}{P(d) \cdot P(q)} = P(d,q) \cdot DQI(d,q) \quad (6)$$

The role of $ES(d,q)$ and $DQI(d,q)$ is explained in Section 5. Mainly, the combination of exhaustiveness and specificity, plus the meaning of DQI, give a meaning to D2Q2.

To estimate $P(q|d)$ and $P(d|q)$, the query q and document d are viewed as *sequences* of independent term events. However, the independence assumption can be seen as sub-optimal. Hence, many approaches such as [Gao *et al.*, 2004; Hou *et al.*, 2011] capture dependence when estimating the document and query probabilities. Similarly, D2Q2 considers dependence by using the notion of semi-subsumed events. The next section reviews this assumption and relates it to the BM25-TF; this justifies why the BM25-TF is later used in D2Q2.

2.4 Semi-subsumed Events

The superior retrieval quality achieved by the BM25-TF is evidence for the dependence of the multiple occurrences of the same term [Robertson *et al.*, 1994]. For instance, [Wu and Roelleke, 2009] pointed out that the BM25-TF

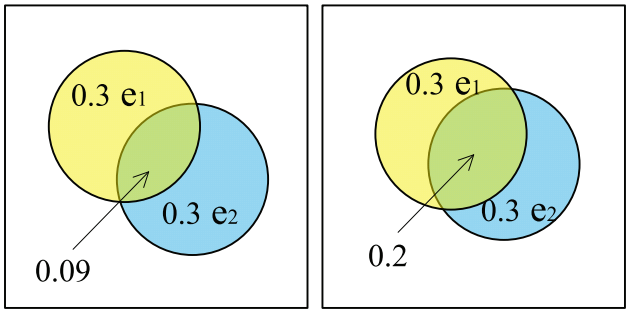

Figure 2: Independent and Semi-subsumed.

can be explained by assuming term occurrences to be *semi-subsumed* events, an important concept for making the proposed hybrid D2Q2 framework a with solid and probabilistic foundations.

In general, the decomposition of event d into term events can be written as:

$$P(d|q) = \prod_{t \in d} P(t|q)^{TF(t,d)} \tag{7}$$

The setting of $TF(t,d)$ reflects probabilistic assumptions:

$$TF(t,d) := \begin{cases} tf_d & \text{independent} \\ 2 \cdot tf_d/(tf_d+1) & \text{semi-subsumed} \\ 1 & \text{subsumed} \end{cases}$$

$TF(t,d) = tf_d$ (total term frequency) views the occurrences as *independent*, whereas $TF(t,d) = 1$ views them as *subsumed* events. Semi-subsumed is between the two. Figure 2 illustrates the computation of $P(e_1, e_2)$ for the case of independent and semi-subsumed events. For IR, event e_i corresponds to the multiple occurrence of a term t_i. For independent events, we obtain $P(e_1, e_2) = 0.3^2 = 0.09$; and for semi-subsumed events, $P(e_1, e_2) = 0.3^{2 \cdot 2/(2+1)} \approx 0.2$. The conjunctive probability of semi-subsumed events is larger than that of independent events. The success of the BM25-TF proves that the multiple occurrences of a term are not independent. The notion of semi-subsumed events assigns a sound semantics to the BM25-TF, making it a well-defined ingredient of D2Q2.

We have discussed the preliminaries of LM and TF-IDF, document-query-(in)dependence (DQI), exhaustiveness and specificity, and semi-subsumed events. The next two sections use these to show the connection between LM and D2, and TF-IDF and Q2.

3 LM as the D2 side of D2Q2

We demonstrate that LM corresponds to the D2 side of D2Q2. We start with reviewing the probabilistic roots of LM as explored in [Hiemstra, 2000; Zhai and Lafferty, 2004]. The notation D2Q stands for $P(q|d)$, and D2 for $P(q|d)/P(q)$, which we denote as D2Q/Q.

$$\text{D2Q} := P(q|d), \qquad \text{D2} := \text{D2Q}\,/Q := \frac{P(q|d)}{P(q)} \tag{8}$$

This section addresses the estimation of $P(q|d)$, or more precisely, of $P(q|d,c)$, where the notation makes explicit the collection "c" used to estimate the background term probability.

3.1 Term (In)dependence Assumption

To estimate $P(q|d,c)$, the query is decomposed into terms:

$$P(q|d,c) = \prod_{t \in q} P(t|d,c)^{TF(t,q)} \tag{9}$$

The conditional d, c makes it explicit that the query and term probabilities depend on both the document d (foreground) *and* the collection c (background). The setting of *TF* reflects two common assumptions made for term events:

$$TF(t,q) := \begin{cases} tf_q & \text{independent} \\ 1 & \text{subsumed} \end{cases} \tag{10}$$

For $P(q|d,c)$, and therefore, D2Q, which assumption is followed is not crucial since often $tf_q = 1$ for short queries. Next we discuss the estimation of $P(t|d,c)$.

3.2 Term Probability Mixture

$P(t|d,c)$ is estimated using a mixture of foreground and background probabilities, essentially to avoid the so-called "zero-probability problem" [Zhai and Lafferty, 2004]. The within-document term probability $P(t|d)$ is mixed with the collection term probability $P(t|c)$ to obtain $P(t|d,c)$:

$$P(t|d,c) = \lambda_d \cdot P(t|d) + (1 - \lambda_d) \cdot P(t|c) \tag{11}$$

The parameter λ_d may be set constant (Jelinek/Mercer mixture, for example, $\lambda_d \approx 0.8$, [Hiemstra, 2000]). Alternatively, $\lambda_d := \frac{dl}{dl+\mu}$ (Dirichlet mixture, *dl* is document length) means that the estimate of $P(t|d)$ is more trusted for longer documents.

We discussed the estimation of $P(q|d,c)$, including the term (in)dependence assumption, leading to the formulation of D2Q. We also referred to D2 as D2Q/Q, that is D2 is equal to D2Q normalised by Q. We discuss the normalisation step next, which leads us to the formulation of D2.

3.3 Normalisation

Applying Equation 9 to Equation 8, making the collection c explicit, and decomposing $P(q|c)$ in the same way as $P(q|d,c)$ (Equation 9), D2 can be decomposed as follows:

$$\text{D2} = \text{D2Q}\,/Q = \frac{P(q|d,c)}{P(q|c)} = \prod_{t \in q} \left(\frac{P(t|d,c)}{P(t|c)} \right)^{TF(t,q)} \tag{12}$$

Using the term probability mixture estimation of $P(t|d,c)$ (Equation 11), we arrive at the following form of D2, which we denote D2-linear, where the subscript indicates the type of the mixture (here a linear mixture):

$$\text{D2}_{\text{linear}} := \prod_{t \in q} \left[(1 - \lambda_d) + \frac{\lambda_d \cdot P(t|d)}{P(t|c)} \right]^{TF(t,q)} \tag{13}$$

We define a second form of D2, denoted D2-extreme, to capture the case of $\lambda_d = 1$ if $t \in d$, and $\lambda_d = 0$ otherwise:

$$\text{D2}_{\text{extreme}} := \prod_{t \in d \cap q} \left[\frac{P(t|d)}{P(t|c)} \right]^{TF(t,q)} \tag{14}$$

We discuss in more detail the extreme mixture when we present Q2, as it establishes the relationship between Q2 and TF-IDF.

3.4 Retrieval Status Value

For each of the D2 forms above, we define an associated retrieval status value (RSV), which can serve as a ranking function. Essentially, the RSV's apply the logarithm.

$$RSV_{\text{D2-linear}}(d,q,c) := \log \text{D2}_{\text{linear}} \tag{15}$$

$$RSV_{\text{D2-extreme}}(d,q,c) := \log \text{D2}_{\text{extreme}} \tag{16}$$

In decomposed form, the RSV's become:

$$RSV_{\text{D2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log\left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)}\right) \tag{17}$$

$$RSV_{\text{D2-extreme}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log \frac{P(t|d)}{P(t|c)} \tag{18}$$

We next make the connection between LM and D2 explicit, namely that D2=LM for the linear form of LM.

3.5 D2 and LM

The following theorem (proof omitted) shows the exact relationship between D2 (the linear form) and LM:

Theorem 1 D2-*linear is an interpretation of LM:*

$$\text{RSV}_{LM}(d,q,c) = \text{RSV}_{\textit{D2-linear}}(d,q,c) \tag{19}$$

Showing that D2=LM does not reveal a new result; the estimation of D2 (leading to D2-linear) was carefully chosen to lead to LM. We nonetheless presented the above steps to prepare for the more complex case demonstrating the relationship between Q2 and TF-IDF.

4 TF-IDF as the Q2 side of D2Q2

We have shown that LM is the D2 $:= P(q|d)/P(q)$ side of D2Q2. Next, we show that TF-IDF is the Q2 $:= P(d|q)/P(d)$ side of D2Q2. This section on TF-IDF is organised analogously to the previous one on LM. For TF-IDF, $P(d|q)$ is the starting point, from where we mirror the steps followed in Section 3. Q2D stands for $P(d|q)$, Q2 for $P(d|q)/P(d)$, where Q2 is Q2D normalised by D denoted Q2D/D.

$$\text{Q2D} := P(d|q), \qquad \text{Q2} := \text{Q2D}\,/D := \frac{P(d|q)}{P(d)} \tag{20}$$

Equation 20 (Q2D) corresponds to Equation 8 (D2Q). Next, we estimate $P(d|q)$.

4.1 Term (In)dependence Assumption

Again we explicate the collection c. To estimate $P(d|q,c)$, the document is decomposed into terms:

$$P(d|q,c) = \prod_{t \in d} P(t|q,c)^{TF(t,d)} \tag{21}$$

Equation 21 corresponds to Equation 9 ($P(q|d,c)$). There are three assumptions encoded in the TF quantification:

$$TF(t,d) := \begin{cases} tf_d & \text{independent} \\ 2 \cdot tf_d/(tf_d + K_d) & \text{semi-subsumed} \\ 1 & \text{subsumed} \end{cases} \tag{22}$$

The semi-subsumed assumption (BM25-TF) led to superior retrieval performance [Robertson *et al.*, 1994]. The parameter K_d is proportional to the pivoted document length $pivdl = dl/avgdl$. The parameter adjusts the semi-subsumption assumption.

4.2 Term Probability Mixture

We again use a mixture model to estimate $P(t|q,c)$:

$$P(t|q,c) = \lambda_q \cdot P(t|q) + (1-\lambda_q) \cdot P(t|c) \tag{23}$$

Equation 23 corresponds to Equation 11 ($P(t|d,c)$).

4.3 Normalisation

Normalisation leads to Q2 (as Q2D/D).

$$\text{Q2} = Q2D/D = \frac{P(d|q,c)}{P(d|c)} = \prod_{t \in d} \left(\frac{P(t|q,c)}{P(t|c)}\right)^{TF(t,d)} \tag{24}$$

Equation 24 corresponds to Equation 12 (D2). As for D2, we define two forms of Q2, linear and extreme. Q2-linear derives directly from applying the term probability mixture to estimate $P(t|q,c)$.

$$\text{Q2}_{\text{linear}} := \prod_{t \in d} \left[(1-\lambda_q) + \frac{\lambda_q \cdot P(t|q)}{P(t|c)}\right]^{TF(t,d)} \tag{25}$$

Equation 25 corresponds to Equation 13 (D2-linear).

The extreme mixture comes from setting $\lambda_q = 1$ if $t \in q$, and $\lambda_q = 0$ otherwise.

$$\text{Q2}_{\text{extreme}} := \prod_{t \in d \cap q} \left(\frac{P(t|q)}{P(t|c)}\right)^{TF(t,d)} \tag{26}$$

Equation 26 corresponds to Equation 14 (D2-extreme). Section 4.6 will show that it is the extreme form of Q2 that is related to TF-IDF.

4.4 Retrieval Status Value

We take the log to define the corresponding retrieval status value for both forms of Q2, and obtain the following:

$$RSV_{\text{Q2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,d) \cdot \log\left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)}\right) \tag{27}$$

$$RSV_{\text{Q2-extreme}}(d,q,c) = \sum_{t \in d \cap q} TF(t,d) \cdot \log \frac{P(t|q)}{P(t|c)} \tag{28}$$

Note the symmetry between Equation 27 and 17, and between Equation 28 and 18.

We continue with Q2-extreme, showing that it corresponds to TF-IDF. Equation 28 has a factor $1/P(t|c)$, the inverse term probability, which reminds of $IDF(t,c) := \log(1/P_D(t|c))$, which we recall is based on the space of *Documents*. However, all the probabilistic estimates so far are based on the space of *Locations* (terms occur at locations). The next section reviews the assumption that allows to transfer the *Location*-based probability $P_L(t|c)$ into the *Document*-based probability $P_D(t|c)$. The transformation between event spaces is necessary to demonstrate since it is one of the pillars between Q2 and TF-IDF.

4.5 Query Term Probability Assumption

We review first the query term probability assumption discussed in [Roelleke and Wang, 2006], which allows the transfer of the Location-based probabilities, $P_L(t|q)/P_L(t|c)$ in Equation 28, to the Document-based probabilities, $1/P_D(t|c)$.

To illustrate the difference between the two spaces, Documents and Locations, consider the following example.

Let term t occur in $tf_c = n_L(t,c) = 1,000$ *Locations* of collection c. Let it occur in $df(t,c) = n_D(t,c) = 200$ *Documents* of collection c. The notation conforms with traditional formulation, and indicates the duality between counting Locations and counting Documents. Then, the average (expected) within-document term frequency is: $avgtf(t,c) = tf_c/df(t,c) = 1,000/200 = 5$. Now let the collection c have $N_L(c) = 10^9$ Locations, and $N_D(c) = 10^6$ Documents. The Location-based probability is $P_L(t|c) = n_L(t,c)/N_L(c) = 1,000/10^9$, the Document-based one is $P_D(t|c) = n_D(t,c)/N_D(c) = 200/10^6$. The average document length is $avgdl(c) = N_L(c)/N_D(c) = 10^3$.

Then, for the fraction of term probabilities, we obtain:

$$\frac{P_L(t|c)}{P_D(t|c)} = \frac{n_L(t,c)/N_L(c)}{n_D(t,c)/N_D(c)} = \frac{avgtf(t,c)}{avgdl(c)} \quad (29)$$

This equation has been referred to as Poisson bridge [Roelleke and Wang, 2006], since it is related to a Poisson probability (we do not need to detail for this paper).

This relationship between Location-based and Document-based term probability enables us to establish the relationship between Q2 and TF-IDF. The relationship is based on the following query term probability assumption:

$$P_L(t|q) = avgtf(t,c)/avgdl(c) \quad (30)$$

What does this assumption express? In the example above, the average document length is $avgdl(c) = 1,000$ and the average within-document term frequency is $avgtf(t,c) = 5$; therefore, $P_L(t|q) = 5/1,000$. With this assumption bursty terms obtain *higher* probabilities than less bursty ones: the query term probability is *proportional* to the burstiness of the term, a reasonable assumption to make.

This assumption leads to $P_L(t|c) = P_L(t|q) \cdot P_D(t|c)$. In turn, this transform the fraction $P_L(t|q)/P_L(t|c)$ (see Equation 28) into an expression based on the Document-based term probability as in IDF:

$$\frac{P_L(t|q)}{P_L(t|c)} = \frac{P_L(t|q)}{P_L(t|q) \cdot P_D(t|c)} = \frac{1}{P_D(t|c)} \quad (31)$$

This establishes the relationship between Q2 and TF-IDF.

$$\log \mathrm{Q2}_{\text{extreme}} = \sum_{t \in d \cap q} TF(t,d) \cdot \log \frac{1}{P_D(t|c)} \quad (32)$$

Next we give the formal proof that shows Q2 (extreme form) is the probabilistic interpretation of TF-IDF.

4.6 Q2 and TF-IDF

In Section 3.5, the relationship between D2 and LM was a direct one. The relationship between Q2 and TF-IDF is less direct, as it relies as above shown on the "query term probability assumption". In addition, whereas showing the relationship between D2 and LM, i.e. LM=D2, relied on a *linear* mixture, showing the relationship between TF-IDF and Q2, i.e. Q2=TF-IDF, relies on the *extreme* mixture.

Given the query term probability assumption, the relationship between Q2 and TF-IDF is expressed as follows.

Theorem 2 $\mathrm{Q2}_{extreme}$ *is an interpretation of* TF-IDF*, if* $P_L(t|q) = P_L(t|c)/P_D(t|c)$*:*

$$P_L(t|q) = \frac{P_L(t|c)}{P_D(t|c)} \Longrightarrow$$
$$\mathrm{RSV}_{TF\text{-}IDF}(d,q,c) = \mathrm{RSV}_{\mathrm{Q2}_{extreme}}(d,q,c) \quad (33)$$

Proof Inserting Equation 3 for $RSV_{\text{TF-IDF}}$ and Equation 28 for $RSV_{\mathrm{Q2}_{\text{extreme}}}$ yields:

$$\sum_t TF(t,d) \cdot TF(t,q) \cdot IDF(t,c) = \sum_t TF(t,d) \cdot \log \frac{P_L(t|q)}{P_L(t|c)}$$

The assumption for $P_L(t|q)$ yields:

$$\frac{P_L(t|q)}{P_L(t|c)} = \frac{P_L(t|c)}{P_D(t|c) \cdot P_L(t|c)} = \frac{1}{P_D(t|c)}$$

Therefore, $\mathrm{Q2}_{\text{extreme}}$ is an interpretation of TF-IDF (for a binary query TF quantification $TF(t,q)$).

We have shown that $\mathrm{D2}_{\text{linear}}$ corresponds to LM, and that $\mathrm{Q2}_{\text{extreme}}$ corresponds to TF-IDF. In the next section, we focus on the relationship between D2 and Q2.

5 On the relationship between D2 (LM) and Q2 (TF-IDF)

Section 2.2 introduced the Document-Query (In)dependence (DQI) measure: $DQI(d,q) := P(d,q)/(P(d) \cdot P(q))$. From the definitions of D2 and Q2, we obtain that $\mathrm{D2} = DQI = \mathrm{Q2}$. This means that D2 and Q2 are equivalent:

$$\mathrm{D2} = \frac{P(q|d,c)}{P(q|c)} = \frac{P(d,q|c)}{P(d|c) \cdot P(q|c)} = \frac{P(d|q,c)}{P(d|c)} = \mathrm{Q2} \quad (34)$$

In other words, before decomposing events into term events and until term (in)dependence assumption made, D2 and Q2 measure the same, that is, LM and TF-IDF aim at measuring the same. The decomposition of d and q into terms breaks the equivalence of D2 and Q2.

$$\sum_{t \in q} TF(t,q) \cdot \log \frac{P(t|d,c)}{P(t|c)} \neq \sum_{t \in d} TF(t,d) \cdot \log \frac{P(t|q,c)}{P(t|c)} \quad (35)$$

For D2, $P(t|d,c)$ is estimated as the linear mixture $\lambda_d \cdot P(t|d) + (1-\lambda_d) \cdot P(t|c)$, establishing that "D2=LM". For Q2, an extreme mixture for $P(t|q,c)$ is applied and we assumed that $P_L(t|q) = P_L(t|c)/P_D(t|c)$, which led to "Q2=TF-IDF".

The following inequality stresses the difference between LM (D2-linear) and TF-IDF (Q2-extreme).

$$TF(t,q) \log \left[(1-\lambda_d) + \lambda_d \cdot \frac{P_L(t|d)}{P_L(t|c)} \right] \neq TF(t,d) \log \frac{1}{P_D(t|c)}$$

We have shown the steps from the equality D2=Q2=DQI that holds before decomposition into term events to the inequality LM≠TF-IDF that comes from the term (in)dependence assumption. This not only shows a relationship between LM and TF-IDF, but explains what connects them, and what separates them.

6 The D2Q2 Framework

We have shown the relationships between LM and D2, between TF-IDF and Q2, between D2 and Q2, and between LM and TF-IDF. The preliminaries introduced concepts (i.e. DQI measure), recalled IR pillars (i.e. exhaustiveness times specificity measure) and relatively recent theory such as semi-subsumed events. Together, the relationships and preliminaries form the theoretical ground of D2Q2.

Our starting point is $ES(d,q) = P(q|d) \cdot P(d|q)$, the exhaustiveness-times-specificity measure commonly used as the basis to justify retrieval models. By analogy, we

define D2Q2 as the product of D2 (Equation 8) and Q2 (Equation 20):

$$\text{D2Q2} := \text{D2} \cdot \text{Q2} \tag{36}$$

where D2 relates to D2Q (exhaustiveness) and Q2 relates to Q2D (specificity). We also know that D2 (linear) corresponds to LM and Q2 (extreme) corresponds to TF-IDF. In other words, D2Q2 "joins" LM and TF-IDF.

We show now that D2Q2 corresponds to DQI^2, where one of the *DQI* relates to LM and the other relates to TF-IDF. This is expressed as follows:

$$\text{D2Q2} = DQI^2$$

By inserting Equation 36 for D2Q2 and Equation 4 for *DQI*, we obtain the decomposed form:

$$\frac{P(q|d)}{P(q)} \cdot \frac{P(d|q)}{P(d)} = \frac{P(d,q)}{P(d) \cdot P(q)} \cdot \frac{P(d,q)}{P(d) \cdot P(q)} \tag{37}$$

We continue now with the two forms of D2Q2, namely, $\text{D2Q2}_{\text{extreme}}$ and $\text{D2Q2}_{\text{linear}}$, which we further decompose:

$$\text{D2Q2}_{\text{extreme}} = \prod_{t \in d \cap q} \left[\left(\frac{P(t|d)}{P(t|c)} \right)^{TF(t,q)} \cdot \left(\frac{P(t|q)}{P(t|c)} \right)^{TF(t,d)} \right] \tag{38}$$

$$\text{D2Q2}_{\text{linear}} = \prod_{t \in d \cap q} \left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)} \right)^{TF(t,q)} \cdot \left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)} \right)^{TF(t,d)} \tag{39}$$

Equations 38 and 39 contain the core contribution of this paper: *the seamless and symmetric composition of probabilistic parameters into a score that embeds LM and TF-IDF*. The main properties of D2Q2 are:

1. A symmetric pattern of the two models' components: for LM these are $P(t|d)$ and $TF(t,q)$, and for TF-IDF these are $P(t|q)$ and $TF(t,d)$; the collection-wide term probability $P(t|c)$ is common to both. The term frequency $TF(t,d)$ and $TF(t,q)$ can be set as in BM25: $TF_K(t,x) := tf_x/(tf_x + K_x)$, which corresponds to assuming the occurrences of t to be semi-subsumed; alternatively, if assuming independence, then $TF(t,x) := tf_x$, where tf_x is the total term frequency count.
2. Derivation and interpretation based on conditional probabilities and document-query independence (*DQI*): $\text{D2} = P(q|d)/P(q) = DQI$ relates to LM, and $\text{Q2} = P(d|q)/P(d) = DQI$ relates to TF-IDF. To decompose D2 and Q2, the "extreme" or the "linear" mixture assumption is applied to both $P(t|d,c)$ and $P(t|q,c)$, leading to $P(t|d)/P(t|c)$ and $P(t|q)/P(t|c)$.
3. The two fractions $P(t|d)/P(t|c)$ and $P(t|q)/P(t|c)$ measure "divergence", i.e. they express that a term with $P(t|d) > P(t|c)$ and $P(t|q) > P(t|c)$ is a good term, where a term is good if its probability in d and q is greater than in collection c. Conditional entropy and Kullback-Leibler divergence incorporate such factors.
4. The "discriminativeness", expressed by $1/P(t|c)$, occurs twice, for the document side and for the query side; this is similar to the vector-space model, where the *idf* is in both the document and query vectors.

For each of D2 and Q2, there is the choice to apply either a linear or the extreme mixture. Our experiments, described next, focus on D2Q2-extreme, which does not involve any mixture parameter, and D2Q2-linear, the model with two mixture parameters (λ_d and λ_q). We define the D2Q2 retrieval status value using logs.

$$RSV_{\text{D2Q2}}(d,q) := \log \text{D2Q2} \tag{40}$$

The next equations show the decomposed, logarithmic form of $\text{D2Q2}_{\text{extreme}}$ (Equation 38) and $\text{D2Q2}_{\text{linear}}$ (Equation 39):

$$RSV_{\text{D2Q2-extreme}}(d,q,c) = \sum_{t \in d \cap q} \left[TF(t,q) \cdot \log \frac{P(t|d)}{P(t|c)} + TF(t,d) \cdot \log \frac{P(t|q)}{P(t|c)} \right] \tag{41}$$

$$RSV_{\text{D2Q2-linear}}(d,q,c) = \sum_{t \in d \cap q} TF(t,q) \cdot \log \left((1-\lambda_d) + \lambda_d \cdot \frac{P(t|d)}{P(t|c)} \right) + TF(t,d) \cdot \log \left((1-\lambda_q) + \lambda_q \cdot \frac{P(t|q)}{P(t|c)} \right) \tag{42}$$

The above decomposed forms illustrates how D2Q2 joins the inner components of LM and TF-IDF, showing that D2Q2 is *hybrid*, i.e. a model beyond combining scores.

7 Experiments

Although the main contribution of this paper was the relationship between LM and TF-IDF, it remains interesting to investigate the experimental performance of D2Q2.

7.1 Set-up

We introduced two retrieval functions derived from D2Q2, $RSV_{\text{D2Q2-extreme}}$ and $RSV_{\text{D2Q2-linear}}$. We now investigate their retrieval performance on a range of collections, outlined in Table 1, of varying size and content.

	Documents $N_D(c)$	Topics $N_Q(c)$	Size
TREC-2	700,000+	50	1.3 GB
TREC-3	700,000+	50	1.3 GB
TREC-8	500,000+	50	834 MB
WT2g	247,000+	50	2 GB
Blogs06	3,200,000+	50	88.8GB

Table 1: Collection Statistics

Following TREC settings [Ounis *et al.*, 2006], for the Blog06 collection, we index only the permalinks (the blog posts and their associated comments). The Porter stemmer was used for stemming. No stopwords removal was applied. We only used the title topic field. We measure retrieval quality with Mean Average Precision (MAP) (topical MAP on Blog06 [Ounis *et al.*, 2006]) and P@10.

Model	Equation
LM_{Dirich}	Equation 2
TF_K-IDF	Equation 3
LM+TF_K-*IDF*	Combinations of retrieval scores
$\text{D2Q2}_{\text{extreme},TF_K}$	Equation 41
$\text{D2Q2}_{\text{linear},TF_K}$	Equation 42

Table 2: Retrieval Models.

Table 2 associates the retrieval models with their respective equations. The first two correspond to the LM and TF-IDF models, the third to the combination of scores of LM

and TF-IDF, and the last two are the two models derived from D2Q2. In TF_K-IDF and in D2Q2, the TF_K component is the BM25-TF, i.e. $TF_K(t,d) = tf_d/(tf_d + K_d)$, where the common setting is $K_d = k_1 \cdot (b \cdot dl/avgdl + (1-b))$. We also set $K_d = 1$ to observe the effect of the BM25-TF on performance.

We used $TF_{\text{K(b=0.25,k1=1.2)}} \cdot$ IDF (which corresponds to BM25 with no relevance information), LM with Dirichlet smoothing and the combination LM+TF_K-*IDF* as baselines. The parameters b, k_1 and μ_D were set to 0.25, 1.2 and 2000, respectively, while μ_Q was set to the average query length. The aforementioned settings were applied across all of the collections, i.e. the retrieval models were not tuned per collection.

For LM+TF_K-*IDF* we used two methods to combine LM and TF_K-IDF inspired by [Larkey and Croft, 1996], and for each method we use two normalisation scheme. The first method is based on *adding* the normalised scores of the documents retrieved by both LM and TF_K-IDF. The normalisation was done either by dividing each individual score by the maximum score for each retrieval model or by dividing by the sum of the scores for each model. The other combination was performed by *multiplying* the normalised scores which were retrieved by both retrieval models. The normalisations were applied in a similar fashion as for the first method.

7.2 Results and Analysis

Table 3 shows for selected models the MAP and P@10. The performance of the TF-IDF with independence assumption, where $TF(t,d) = tf_d$, was omitted since too poor to be considered as a baseline (MAP in average was one third of the MAP achieved by TF_K-IDF). Similar observations were made for D2Q2 with independence assumption, and as such the corresponding results are omitted.[2]

The setting $TF_K := tf/(tf + K)$ was instrumental in achieving competitive retrieval performance, and hence we report only results for this setting. We discussed the notion of "semi-subsumed" events which embeds the BM25-TF into D2Q2. In D2, $TF_K(t,q)$ is applied whereas in Q2, it is $TF_K(t,d)$. D2Q2-extreme has no mixture parameters, whereas for D2Q2-linear, the parameter μ_D controls the Dirichlet mixture parameter λ_d (and μ_Q controls λ_q). The overall result is expressed by the relative distance between models (last row of Table 3).

Overall, most candidates deliver about the same performance, with marginal differences among the top candidates. Only one score combination (multiplication of normalised LM and TF-IDF scores) is a poor outlier. The D2Q2 family of models has in half of the cases (5 of 10 benchmarks) the best performer. Some members of the D2Q2 family performed better than others, where in tendency, the linear mixtures are better than extreme mixtures. This is as expected, since the extreme mixtures rely on assumptions that neglect the Dirichlet mixture parameter.

We ran statistical significance tests based on Student's paired t-test with confidence levels $\alpha = 0.01$ and $\alpha = 0.05$. In all cases, the results for the best D2Q2 model and the best traditional model were not significantly different. On one hand, this confirms the reasonable performance of the D2Q2 models. On the other hand, if we had expected an improvement from devising a new model that consists of the inner organs of LM and TF-IDF, then we are disappointed, since the single models perform already relatively well on their own.

Overall, the experimental results show that the hybrid D2Q2 performs within the main-fold of the retrieval quality reported for the baselines. Regarding the comparison of the *score aggregation* LM+TF_K-*IDF* versus the *hybrid* D2Q2, the score aggregation is outperformed by the hybrid (except for TREC-2 where the difference is marginal). In the light of the aforementioned expectation that combining two models delivers the averaged quality, the performance of D2Q2 underlines the effect of hybridity. This supports the conclusion that D2Q2 combines the LM and TF-IDF features such that a *micro* combination of probabilities performs better than a *macro* combination of scores as expressed by LM+TF_K-*IDF*.

D2Q2 shows a stable performance that is marginally better than the baselines, but D2Q2 does not significantly outperform the baselines. The experiments confirm the rationale underlying D2Q2, a framework that encompasses LM and TF-IDF, and their combinations. In particular, D2Q2 truly combines the LM and TF-IDF features into a theory based on probabilities, exhaustiveness and specificity.

8 Conclusions

This research was motivated by investigating the relationship between LM and TF-IDF to attempt to provide answers to statements such as "we know *why* TF-IDF works, and we know *that* LM works, but we do not know *why* LM works". By developing a side-by-side derivation of LM and TF-IDF, a framework based on $P(q|d) \cdot P(d|q)$ emerged, which we named D2Q2. The main contribution of this paper is the theory that underpins the probabilistic framework D2Q2, where the D2 side is LM, and the Q2 side is TF-IDF. This theory reveals the link between LM and TF-IDF, and the D2Q2 framework shows how the features of both models can be combined in a theoretically sound manner. In addition, D2Q2 shows comparable retrieval performance to competitive baselines, making D2Q2 to be not just another unifying framework but a retrieval model in its own right.

Our emphasis was on LM and TF-IDF. Future work will elaborate on the relationship between BM25 and D2Q2. D2Q2 establishes a balanced view on LM and TF-IDF, and this can potentially lead to a consolidated anatomy of the models, viewing LM and TF-IDF as the models for missing relevance, and devising BM25-D2 (an LM-based BM25) and BM25-Q2 (TF-IDF-BM25) as relevance models.

References

[Aizawa, 2003] Akiko Aizawa. An information-theoretic perspective of TF-IDF measures. *IP&M*, 39:45–65, 2003.

[Bartell *et al.*, 1994] B. T. Bartell, G. W. Cottrell, and R. K. Belew. Automatic combination of multiple ranked retrieval systems. *ACM SIGIR*, pages 173–181, 1994.

[Church and Gale, 1995a] K. Church and W. Gale. Inverse document frequency (IDF): A measure of deviation from Poisson. *Workshop on Very Large Corpora*, pages 121–130, 1995.

[Church and Gale, 1995b] K. Church and W. Gale. Poisson mixture. *Natural Language Engineering*, 1(2):163–190, 1995.

[Croft *et al.*, 1990] W.B. Croft, R. Krovetz, and H. Turtle. Interactive retrieval of complex documents. *IP&M*, 26(5):593–613, 1990.

[2] We could however notice that the independence assumption was less detrimental for the D2 (LM) side than for the Q2 (TF-IDF) side. This is because for D2, the assumption is for the query ($TF(t,q)$), which usually contains only few multiple occurrences of terms.

	TREC-2		TREC-3		TREC-8		WT2g		Blog06	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
$LM_{\text{Dir},\mu=2000}$	18.02	41.20	22.87	48.20	21.48	40.00	29.85	46.20	29.21	60.80
$TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$	**18.90**	*42.80*	**25.05**	*50.20*	22.31	40.20	*31.42*	**49.20**	*30.27*	**63.40**
LM+TF-IDF										
$LM_{\text{Dir},\mu=2000} + TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Max norm)	18.65	**44.40**	24.46	49.20	*22.45*	*41.40*	31.15	47.40	29.56	61.20
$LM_{\text{Dir},\mu=2000} + TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Sum norm)	18.72	44.00	24.30	49.60	22.38	*41.40*	31.16	47.20	29.85	62.80
$LM_{\text{Dir},\mu=2000} \cdot TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Max norm)	13.56	43.20	18.80	47.60	19.35	41.00	26.93	47.00	27.10	60.00
$LM_{\text{Dir},\mu=2000} \cdot TF_{\text{K(b=0.25,k1=1.2)}} \cdot \text{IDF}$ (Sum norm)	6.58	5.80	6.39	5.20	6.65	4.60	6.54	5.60	21.97	32.80
D2Q2										
$\text{D2Q2}_{\text{extreme},TF_{\text{K=1}},TF(t,q)=1}$	17.59	42.80	23.00	47.00	**23.16**	42.40	31.92	45.20	29.44	55.80
$\text{D2Q2}_{\text{extreme},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=0.5}$	16.89	38.40	20.73	40.20	22.08	42.00	31.74	46.40	29.22	58.00
$\text{D2Q2}_{\text{extreme},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1}$	17.24	40.00	24.17	49.80	22.65	**43.60**	28.80	44.80	28.23	54.60
$\text{D2Q2}_{\text{linear},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1,\mu_D=2000}$	18.48	44.00	24.81	**51.00**	22.52	42.20	31.36	48.00	29.85	62.40
$\text{D2Q2}_{\text{linear},TF_{\text{K(b=0.25,k1=1.2)}},TF(t,q)=1,\mu_D=2000,\mu_Q=avgql}$	18.60	44.20	24.83	50.40	22.59	42.20	31.38	48.00	29.87	62.20
$\text{D2Q2}_{\text{linear},TF_{\text{K=1}},TF(t,q)=1,\mu_D=2000,\mu_Q=avgql}$	17.13	41.20	20.81	40.20	21.68	40.60	**32.03**	46.60	**30.40**	61.80
best D2Q2 - best traditional	-0.30	-0.20	-0.22	0.80	0.71	2.20	0.61	-1.20	0.13	-1.00
relative difference	-0.01	-0.004	-0.008	0.01	0.03	0.05	0.02	-0.02	0.004	-0.01

Table 3: MAP and P@10 (Best traditional model *italicised*, best overall model in **bold).**

[Croft, 2000] W. Bruce Croft. *Combining Approaches to Information Retrieval*, IR 1, pages 01–36. Kluwer Academic Publishers, 2000.

[Fang and Zhai, 2005] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. *ACM SIGIR*, pages 480–487, 2005.

[Frei *et al.*, 1996] H.P. Frei, D. Harmann, P. Schäuble, and R. Wilkinson, editors. *ACM SIGIR*, 1996.

[Gale and Church, 1991] William A. Gale and Kenneth Ward Church. Identifying word correspondences in parallel texts. *HLT*, 1991.

[Gao *et al.*, 2004] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. *ACM SIGIR*, pages 170–177, 2004.

[He and Ounis, 2005] Ben He and Iadh Ounis. A study of the dirichlet priors for term frequency normalisation. *ACM SIGIR*, pages 465–471, 2005.

[Hiemstra, 2000] Djoerd Hiemstra. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[Hou *et al.*, 2011] Y. Hou, L. He, X. Zhao, and D. Song. Pure high-order word dependence mining via information geometry. *Advances in Information Retrieval Theory ICTIR*, pages 64–76. 2011.

[Kwok, 1996] K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. I*ACM SIGIR*, pages 187–195, 1996.

[Lafferty and Zhai, 2003] John Lafferty and ChengXiang Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, chapter 1. Kluwer, 2003.

[Larkey and Croft, 1996] L. S. Larkey and W. B. Croft. Combining classifiers in text categorization. [1996], pages 289–297.

[Lee,] Joon Ho Lee. Analyses of multiple evidence combination. *SIGIR Forum*, 31(SI):267–276.

[Metzler and Croft, 2004] Donald Metzler and W. Bruce Croft. Combining the language model and inference network approaches to retrieval. *IP&M*, 40(5):735–750, 2004.

[Nie, 1992] J.J. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. *ACM SIGIR*, pages 140–151, 1992.

[Ounis *et al.*, 2006] Iadh Ounis, Craig Macdonald, Maarten de Rijke, Gilad Mishne, and Ian Soboroff. Overview of the trec 2006 blog track. *TREC*, 2006.

[Ponte and Croft, 1998] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. *ACM SIGIR*, pages 275–281, 1998.

[Robertson *et al.*, 1994] S. Robertson, S. Jones S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. *Text REtrieval Conference*, 1994.

[Robertson, 2004] S.E. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.

[Roelleke and Wang, 2006] Thomas Roelleke and Jun Wang. A parallel derivation of probabilistic information retrieval models. In *ACM SIGIR*, pages 107–114, 2006.

[Roelleke and Wang, 2008] Thomas Roelleke and Jun Wang. TF-IDF uncovered: A study of theories and probabilities. In *ACM SIGIR*, pages 435–442, 2008.

[Salton *et al.*, 1976] G. Salton, A. Wong, and C.T. Yu. Automatic indexing using term discrimination and term precision. *IP&M*, 12:43–56, 1976.

[Singhal *et al.*, 1996] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalisation. [1996], pages 21–39.

[Taylor *et al.*, 2006] Michael Taylor, Hugo Zaragoza, Nick Craswell, Stephen Robertson, and Chris Burges. Optimisation methods for ranking functions with multiple parameters. *ACM CIKM*, 2006.

[Wong and Yao, 1995] S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst.*, 13(1):38–68, 1995.

[Wu and Roelleke, 2009] Hengzhi Wu and Thomas Roelleke. Semi-subsumed events: A probabilistic semantics for the BM25 term frequency quantification. *ICTIR (International Conference on Theory in Information Retrieval)*, 2009.

[Wu *et al.*, 2008] Ho Chung Wu, Robert Wing Pong Luk, Kam-Fai Wong, and Kui-Lam Kwok. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26(3), 2008.

[Zaragoza *et al.*, 2003] Hugo Zaragoza, Djoerd Hiemstra, and Michael Tipping. Bayesian extension to the language model for ad hoc information retrieval. *ACM SIGIR*, pages 4–9, 2003.

[Zhai and Lafferty, 2001] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *ACM SIGIR*, pages 334–342, 2001.

[Zhai and Lafferty, 2004] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.