Interpreting TF-IDF Term Weights as Making Relevance Decisions

HO CHUNG WU and ROBERT WING PONG LUK
The Hong Kong Polytechnic University
KAM FAI WONG
The Chinese University of Hong Kong
and
KUI LAM KWOK
Queens College, City University of New York

A novel probabilistic retrieval model is presented. It forms a basis to interpret the TF-IDF term weights as making relevance decisions. It simulates the local relevance decision-making for every location of a document, and combines all of these "local" relevance decisions as the "document-wide" relevance decision for the document. The significance of interpreting TF-IDF in this way is the potential to: (1) establish a unifying perspective about information retrieval as relevance decision-making; and (2) develop advanced TF-IDF-related term weights for future elaborate retrieval models. Our novel retrieval model is simplified to a basic ranking formula that directly corresponds to the TF-IDF term weights. In general, we show that the term-frequency factor of the ranking formula can be rendered into different term-frequency factors of existing retrieval systems. In the basic ranking formula, the remaining quantity $-\log p(\bar{r}|t \in d)$ is interpreted as the probability of randomly picking a nonrelevant usage (denoted by $\bar{r}$) of term t. Mathematically, we show that this quantity can be approximated by the inverse document-frequency (IDF). Empirically, we show that this quantity is related to IDF, using four reference TREC ad hoc retrieval data collections.

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

General Terms: Design, Experimentation, Languages, Performance

Additional Key Words and Phrases: Information retrieval, term weight, relevance decision

This research was supported by the CERG Project no. PolyU 5226/05E.

Authors' addresses: H. C. Wu, R. W. P. Luk (corresponding author), Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong; email: csrluk@comp.polyu.edu.hk; K. F. Wong, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, The People's Republic of China; K. L. Kwok Department of Computer Science, Queens College, City University of New York, 65-30 Kissena Blvd., Flushing, NY 11367.

ACM Reference Format:
Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. 2008. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inform. Syst. 26, 3, Article 13 (June 2008), 37 pages. DOI = 10.1145/1361684.1361686 http://doi.acm.org/ 10.1145/1361684.1361686

1. INTRODUCTION

This article presents a basis to interpret the well-known TF-IDF term weights [Luhn 1958; Robertson and Spärck Jones 1976; Yu and Salton 1976; Amati and van Rijsbergen 2002] as making relevance decisions. This basis is our novel probabilistic retrieval model that simulates human relevance decision-making for two types of relevance. One new type is the "local" relevance that only applies to a specific document location, and the other, common, type is the "document-wide" relevance that applies to the entire document. The model combines the local relevance for every location of a document by the document-wide relevance decision of the document. The local relevance at location k is the outcome of the local relevance decision, which is made on the basis of the available information in the document context centered at k. If the document is locally relevant at any document location, then the entire document will be deemed document-wide relevant to the query. This way of combining local relevance at different document locations to arrive at a document-wide relevance decision is consistent with the TREC ad hoc evaluation policy [Harman 2004] as described in Section 2.4. This policy is also applied to terabyte ad hoc retrieval (e.g., Clarke et al. [2005]), multimedia retrieval (e.g., Clough et al. [2004]), and XML element retrieval (e.g., Trotman and Geva [2006]).

We are motivated to justify mathematically and empirically that TF-IDF term weights can be the outcome of modeling relevance decision-making. The significance of this justification is the potential that there is a unifying perspective about information retrieval (IR) as relevance decision-making. Many past retrieval models are already related to relevance decision-making; for example, the binary independence model [Robertson and Spärck Jones 1976], logistic regression model [Cooper et al. 1992], vector space model [Salton et al. 1975], Boolean model [Wong et al. 1986], and extended Boolean model [Salton et al. 1983]. However, it is not known whether TF-IDF term weights are related to relevance decision-making because they were originally not conceived in this way. Instead, the term-frequency factor was originally thought to be indicative of document topic [Luhn 1958], and the inverse document-frequency (IDF) is reasoned [Spärck Jones 1972] on the basis of Zipf law.

The original TF-IDF term weights are thought to be attribute values of documents that are treated as indivisible objects in many IR models. From our novel perspective, TF-IDF term weights are treated as the outcome of local relevance decision-making at different document locations. This novel perspective is a new avenue to develop more novel retrieval models, and it extends the original TF-IDF term weights to model microscopic phenomena at the document-location level, rather than macroscopic phenomena at the document level. This new perspective also demands a new representation of a document

as a string of words, instead of the common vector representation, because the string representation of a document exposes information in the document for the purpose of mathematical modeling.

The simplified basic ranking formula of our probabilistic retrieval model that provides a basis to interpret TF-IDF term weight is the probability of relevance $p(R_{d,q} = r)$ that is rank equivalent (i.e., denoted by $\propto$) to the sum of products

$$p(R_{d,q} = r) \propto \sum_{t \in (V(q) \cap V(d))} f(t, d) \times [-\log p(\bar{r}|t \in d)],$$

where $f(t, d)$ is the occurrence frequency of t in d and the quantity $-\log p(\bar{r}|t \in d)$ corresponds to IDF. Details of the symbols and their descriptions of the previous formula are listed in Table I.

The previous basic ranking formula is consistent with the probability ranking principle [Robertson 1977] because it ranks documents by probability of relevance. The term-frequency factor of the basic ranking formula is $f(t, d)$, and the remaining quantity $-\log p(\bar{r}|t \in d)$ is approximated by IDF. This approximation of $-\log p(\bar{r}|t \in d)$ is also supported empirically, using four reference TREC ad hoc test collections. For generality of modeling, the quantity $-\log p(\bar{r}|t \in d)$ can also be approximated by the inverse collection term-frequency (ICTF) [Kwok 1995], which has been found to correlate with IDF using those reference ad hoc test collections. An independent, empirical approach, using clustering to estimate the quantity $-\log p(\bar{r}|t \in d)$, illustrates the explanatory power of the previous basic ranking formula.

The rest of this article is organized as follows. Section 2 describes our novel probabilistic retrieval model that forms a basis to interpret TF-IDF term weights. The ranking formula of this model is simplified to the basic ranking formula that directly corresponds to TF-IDF term weights. Section 3 interprets the quantity $p(\bar{r}|t \in d)$ of the basic ranking formula as the probability of randomly picking a nonrelevant usage of term t. We show that $-\log p(\bar{r}|t \in d)$ can be approximated by IDF. Another independent, empirical approach directly estimates $p(\bar{r}|t \in d)$ using a novel clustering algorithm. Section 4 reports on the experiments relating to this approach. Section 5 describes related work. Section 6 concludes this work.

2. PROBABILISTIC NONRELEVANCE DECISION MODEL

We formulate our probabilistic model as follows. Section 2.1 specifies the general model using document-context ranking, and it justifies the use of document contexts. Section 2.2 describes the probability notation used in this article. Section 2.3 develops our probabilistic model and derives the context-based ranking formula (Eq. (6)). In Section 2.4, this formula is simplified to the basic ranking formula that directly corresponds to the TF-IDF term weights.

2.1 General Model

Recently, Wu et al. [2007] proposed a novel retrieval model that achieved high retrieval effectiveness (i.e., between 60% and 80% mean average precision) in their retrospective experiments using several ad hoc test collections. The authors implicitly distinguish two types of relevance: the common document-wide

Table I. Mathematical Symbols Used and Their Descriptions

Symbols	Description
D	A collection of documents
Q	A set of queries
R	A set of relevance values (r for relevance, $\bar{r}$ for irrelevance)
U	A set of human evaluators
$card(.)$	The cardinality of its argument
d	A document (typically considered as a string or set of words)
$\|d\|$	The length (total number of terms) of the document d
$d[k]$	The term located at the kth logical position in document d
$c(d, k, n)$	A context of size $2n + 1$ terms located at position k in document d
q	A query
$\propto$	Rank-equivalence binary relation
$\equiv$	Defined as or assignment relation
$\nabla(d, q)$	Document-wide relevance-decision function for document d and query q
$\partial_{d,k}(c(d, k, n), q)$	Local relevance-decision function at location k in document d for query q
$C(.)$	The generic function that combines the outcomes of local relevance decisions
N	The set of positive integers
Ω	Event space
Ω_{∇}	Event space for document-wide relevance decision-making
$\Omega_{\partial,n}$	Event space for local relevance decision using context of size $2n + 1$
$\succcurlyeq$	The weak ordering relation of document-wide or local decision preferences
$f(t, d)$	The occurrence frequency of term t in document d
$tf(t)$	The total occurrence frequency of term t in all documents
$tf(D)$	The sum of occurrence frequencies of all terms in the collection D
$rtf(t,q)$	The total occurrence frequency of t in documents relevant to q
$avgtf(D)$	The average number of terms in a document of collection D
$Loc(t, d)$	The set of locations of term t in document d
$df(t)$	The document frequency of t
$rdf(t, q)$	The number of documents that contain term t and relevant to q
$IDF(t)$	Inverse document-frequency of term t
$QIDF(t, q)$	Query-dependent inverse document-frequency of term t
$p_{\Omega}(.)$	The probability of which argument is in the event space Ω
$\wp(.)$	The power set of its argument
$v(.)$	The vector representation of its argument
$\Rightarrow$	One-way implication
$\Longleftrightarrow$	Two-way implication
$\wedge$	Conjunction operator
$R_{d,q}$	Document-wide relevance variable for document d and query q
$R_{d,k,q}$	Local relevance variable at location k of document d for query q
α	A linear mixture parameter
β	The scaling parameter in the systematic sampling
$B(t)$	The set of unique usages of t
$M(t)$	The total number of usages of the term t
$\Lambda(t)$	The arrival rate of the new usages of t per document
$\Lambda_c(t)$	The arrival rate of the new usages of t per occurrence
$m(t)$	The number of new usages of t for the random match model
$\eta(t)$	The Poisson distributed random variable representing the number of new usages of t
$\bullet$	The dot product of two vectors
$\alpha(t)$	The number of usages of the term t which are relevant to any query
$h(t, q)$	The number of unique relevant usages of the term t to the query q
$W_E(t)$	The expectation weight of term t
$C_E(t)$	The context-counting expectation weight of term t
$E(.)$	The expectation operator
$V(.)$	The set of distinct terms of its argument

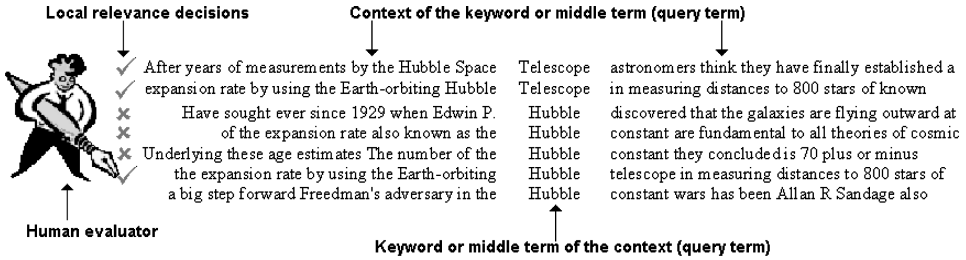

Fig. 1. A qualitative model of the process of making human judgment based on combining the local relevance decisions, as in Wu et al. [2007]. Example contexts are extracted from a document (NYT19990525.0358) relevant to query 303 in the TREC-2005 robust-track data collection. The query is "Hubble Telescope Achievements". Contexts with (✓) are judged relevant by a human evaluator.

relevance $R_{d,q}$ that applies to the entire document d for query q, and the new local relevance $R_{d,k,q}$ that applies only at the document location k in d for q. Both local and document-wide relevance can be binary values (e.g., $\overline{r}$ or r, or numerically, 0 or 1), or real values representing the degree of local and document-wide relevance, respectively. Typically, these real values are normalized between 0 and 1, without loss of generality.

The model by Wu et al. [2007] simulates a human evaluator who scans the document for local relevance information (see Figure 1). Scanning involves iterating through every document location, and deciding for each whether local relevance information is found. The local relevance for each document location is combined to form the document-wide relevance of the entire document. Mathematically, the document-wide relevance is specified by the general equation

$$R_{d,q} \equiv C(\{R_{d,k,q} : k \in [1, |d|], k \in N\}), \tag{1}$$

where N is the set of positive integers, $|d|$ is the length of document d, and $C(.)$ is the general mathematical function that combines the local relevance. In this article, we assume that the first location of any document starts at 1.

According to Wu et al. [2007], the outcome of a local relevance decision at location k of document d is determined by the information in the context that is denoted by $c(d, k, n)$. This context has n terms on the left, and another n terms on the right from location k in d. Figure 1 shows examples of information extracted as contexts from a document for the query "Hubble Space Telescope". The keyword or middle term of the context is the term being scanned at present, and the human evaluator decides whether the information in the context of the middle term is locally relevant at that location.

The use of document context assumes that document information that is far away from location k has negligible impact on the local relevance decision at location k. This is supported by past studies which found that: (1) the n-dependence entropy asymptotically approaches towards the entropy of a random model of character sequences [Wong and Ghahraman 1975]; and (2) the mutual information of English text [Lucassen and Mercer 1984] and Chinese text [Hung et al. 2001] decreases as the distance increases between the term in the middle of the context and the other term in the context. In local context

analysis (LCA) [Xu and Croft 2000] or lexical cohesion [Vechtomova et al. 2006], it is implicitly assumed that terms far away from other terms in the middle of the context have negligible impact, and thus such terms are ignored in the LCA and lexical cohesion. Wu et al. [2007] also showed their retrieval model to obtain high retrieval effectiveness in retrospective experiments, using a fixed-size context that is much shorter than the document length. Their work [Wu et al. 2006, 2005] in relevance feedback using a document context model for reranking documents also reported effective retrieval. Their work directly supports the use of document contexts for local relevance decision-making.

After defining document contexts and supporting their use in information retrieval, we assume the following to simplify the modeling of local relevance decision-making.

Assumption (*Context-Based Local Relevance Decision*). A local relevance decision at any location k in any document d for any query q is made on the basis of the information in the context that is centered at k in d for some maximal context size n.

To model relevance decisions, we denote $\partial_{d,k}(.)$ as the local relevance decision at location k of document d. It is location based because local relevance is location dependent. According to the previous assumption, the input of local relevance decisions consists of the context $c(d, k, n)$ and the query q. Its output is the decision preference (as in Yao et al. [1991]) of the local relevance assigned by the human evaluator. According to ordinal value theory [French 1986, Chapter 3], this decision preference can be transformed into a real value in [0,1]. For notational simplicity, we assume that $\partial_{d,k}(.)$ produces such an ordinal value that represents the local relevance decision preference. Therefore, the local relevance $R_{d,k,q}$ at k in d for q is the outcome of the corresponding local relevance decision at k in d, as follows.

$$R_{d,k,q} \equiv \partial_{d,k}(c(d, k, n), q)$$

If $\partial_{d,k}(.)$ only returns 0 for local nonrelevance and 1 for local relevance, then $R_{d,k,q}$ will be a binary variable for local relevance. Although $\partial_{d,k}(.)$ can be a real value in [0,1], we restrict our discussion in this article to binary variables for simplicity and clarity of representation. Similar to the local relevance variable, we assign the document-wide relevance variable $R_{d,q}$ with $\nabla(d, q)$ that contains the binary, document-wide relevance.

$$R_{d,q} \equiv \nabla(d, q)$$

Using the definitions of local- and document-wide relevance, Eq. (1) is specified in terms of making relevance decisions, as follows.

$$\nabla(d, q) = C(\{\partial_{d,k}(c(d, k, n), q) : k \in [1, |d|], k \in Z\}) \tag{2}$$

The previous equation provides a direct, general mathematical description of the human evaluator making relevance decisions using a document-context-based model for local relevance decision-making. It generalizes recent work by Wu et al. [2007, 2006, 2005]. Their work modeled the set of local relevance, $\{R_{d,k,q}\}$, as local decision preferences that are defined as the normalized log-odds

[Robertson and Spärck Jones 1976] of the local relevance of the corresponding document contexts $\{c(d, k, n)\}$. Wu et al. [2007] experimented with several different implementations of the combining function $C(.)$ (e.g., the extended Boolean disjunction [Salton et al. 1983], fuzzy disjunction [Dombi 1982], or order-weighted averaging operators [Waller and Kraft 1979; Paice 1984]). In contrast, $R_{d,q}$ and $R_{d,k,q}$ in this article are formulated as (random) binary variables in our probabilistic formulation for binary relevance (discussed in the next section). Instead of determining the output of the local relevance decisions, our probabilistic formulation combines the probability of local relevance decisions with the desirable outcomes to estimate the probability of document-wide relevance.

2.2 Probability Notation and Interpretation

To specify the event space [Robertson 2005] of relevance decisions, let us denote the following. Let:

—the set of documents be D;
—the set of queries be Q;
—the set of relevance values be R (i.e., for binary relevance, $R = \{r, \bar{r}\}$);
—the set of human evaluators be U;
—the set of terms, in the collection D, be $V(D)$;
—the set of possible strings, of length $2n + 1$ over $V(D)$, be $V(D)^{2n+1}$, where $V(D)^{2n+1}$ is the cross-product of $V(D)$ itself by $2n + 1$ times.

The event space Ω_∇ of document-wide relevance decisions is defined as

$$\Omega_\nabla = (D \times Q \times R)^{card(U)}.$$

This event space Ω_∇ represents an experiment that is repeated for the set U of human evaluators who are asked to assign their decision preferences concerning a document in D for a given query in Q. The event space $\Omega_{\partial,n}$ of the local relevance decisions is defined only as a subset of the cross-product space

$$\Omega_{\partial,n} \subseteq \left(D \times N \times V(D)^{2n+1} \times Q \times R\right)^{card(U)}$$

because certain events in the cross-product space are undefined in $\Omega_{\partial,n}$ (e.g., the integer value is larger than the longest document in the collection). This event space is specified by a context-size parameter n. The context string $c(d, k, n)$, of length $2n + 1$, is specified over $V(D)^{2n+1}$. This context can be deduced from d and k, but is explicitly modeled because it corresponds to the document context model of Wu et al. [2007]. When k is near the beginning or end of the document, the context string is padded with a special unique character to ensure that the length of the string is $2n + 1$. We assume that this special character is already in $V(D)$. The set R is the set of possible values for the local relevance. The aforementioned cross-product space includes the set N of positive integers, because the events are specified by the document locations. The set R in Ω_∇ is different from the one in $\Omega_{\partial,n}$ because R in Ω_∇ is comprised of values taken

by the document-wide relevance variable, whereas R in $\Omega_{\partial,n}$ consists of values taken by the local relevance variable.

We denote the probability of document-wide relevance as $p_\nabla(R_{d,q})$, where the subscript ∇ identifies the event space as Ω_∇, and ∇ specifies that the relevance value of $R_{d,q}$ is produced by the document-wide relevance decision $\nabla(.)$. Detailed arguments for $\nabla(.)$ are not necessary because they are completely specified by $R_{d,q}$ according to its definition. Similarly, the probability of local relevance is denoted by $p_{\partial,n}(R_{d,k,q})$, where ∂ identifies the event space as Ω_∂, and it specifies that the local relevance value of $R_{d,k,q}$ is produced by the local relevance decision $\partial(.)$, with context size $2n + 1$. Detailed subscripts and arguments for $\partial(.)$ are not necessary because they are completely specified by $R_{d,k,q}$ according to its definition, apart from the context size n.

When $card(U) > 1$, the probabilities are interpreted as the proportion of human evaluators. If these evaluators make relevance decisions independent of each other, then the experiment will be treated as a $card(U)$ number of Bernoulli trials. When $card(U) = 1$, the probabilities are interpreted as the assigned degree of belief that the information item is relevant to the query. The degrees of belief are treated as decision preferences [Yao et al. 1991] about the relevance of the information. Such a preference value need not be the same as the degree of belief of relevance. Ideally, the weak ordering of the degree of belief of the local- and document-wide relevance is the same as the weak ordering of the corresponding decision preferences for local- and document-wide relevance, respectively. Formally, for query q, for any two documents d and e, and for any valid document locations j and k, it holds that

$$p_{\partial,n}(R_{d,j,q} = r) \geq p_{\partial,n}(R_{e,k,q} = r) \Longleftrightarrow \partial_{d,j}(c(d, j, n), q) \succcurlyeq \partial_{e,k}(c(e, k, n), q)$$

for local relevance. Moreover, for q, and for any d and any e in D, it holds that

$$p_\nabla(R_{d,q} = r) \geq p_\nabla(R_{e,q} = r) \Longleftrightarrow \nabla(d, q) \succcurlyeq \nabla(e, q)$$

for document-wide relevance. In practice, this perfect representation of the decision preferences by the degrees of belief seldom occurs, so the retrieval effectiveness is seldom perfect (i.e., 100% mean average precision). For the purpose of modeling, we assume that the degree of belief represents the decision preferences, and the remaining practical problem of modeling is to estimate this degree of belief from data, without the need to manually assign it by the human evaluator.

2.3 Context-Based Ranking Formula

We model the relevance decision with nonrelevance outcomes (similar to Calado et al. [2003]), and we rank documents by the probability of nonrelevance in reverse order. For binary relevance, $p_\nabla(R_{d,q} = r)$ can be expressed as

$$p_\nabla(R_{d,q} = r) = 1 - p_\nabla(R_{d,q} = \overline{r}). \tag{3}$$

The probability of document-wide nonrelevance in Eq. (3) can be expressed in terms of the probability of local nonrelevance by using the TREC evaluation policy for ad hoc retrieval tasks. According to Harman [2004], if any part of a document is judged relevant to the topic, then the entire document is considered

as relevant in a TREC ad hoc retrieval task. Such an evaluation policy for ad hoc retrieval is used because ad hoc retrieval tasks are supposed to be recall oriented, and because such an inclusive policy enables later research on more specific relevance judgments [Harman 2004]. Given this understanding of the evaluation policy and that we are dealing with binary relevance, a document d will be deemed document-wide not relevant to a query if every local relevance decision in the document is not relevant.

Logically, the TREC evaluation policy for ad hoc retrieval tasks is specified as a two-way implication as

$$(R_{d,q} = \overline{r}) \Longleftrightarrow \bigwedge_{k=1}^{|d|} (R_{d,k,q} = \overline{r}),$$

where $=$ is the equality test that returns true if the values are the same, and false otherwise. The previous logical relationship is a Boolean logic version of Eq. (1), where $C(.)$ in (1) is specified as a conjunction in Boolean logic. Based on this logical relationship and Eq. (2), the probability that the document is not relevant can be assigned as the joint probability that all local relevance of individual document locations is not relevant.

$$p_{\nabla}(R_{d,q} = \overline{r}) \equiv p_{\partial,n}((R_{d,1,q} = \overline{r}), \ldots, (R_{d,|d|,q} = \overline{r}))$$

Note that the event spaces on the lefthand side (LHS) and righthand side (RHS) of the previous equation are different. This is because the equation relates the two types of relevance, the document-wide- and local relevance. From the perspective of mathematical modeling, the joint probability on the RHS of the previous equation simulates the local relevance decision-making with nonrelevance outcome for the document d. The estimated joint probability is assigned to the probability of document-wide nonrelevance on the LHS. It is expected in mathematical modeling that this probability assignment (on the RHS) is unlikely to be exactly the same as the true probability (on the LHS) because we do not expect perfect retrieval effectiveness performance. The question is whether the error of this probability assignment will have some impact on the retrieval effectiveness. To reduce this impact of error and yet without loss of generality, the assigned probability (on the RHS) is made rank equivalent with the true probability (on the LHS). Using this rank equivalence relation, Eq. (3) becomes

$$p_{\nabla}(R_{d,q} = r) \propto -\log p_{\partial,n}((R_{d,1,q} = \overline{r}), \ldots, (R_{d,|d|,q} = \overline{r})). \tag{4}$$

In order to simplify the previous equation, we assume that the local relevance decisions with nonrelevance outcomes are independent. Specifically, we give the next assumption.

Assumption (*Nonrelevance Independence*). For any document d and any query q, $p_{\partial,n}(R_{d,k,q} = \overline{r} | R_{d,k-1,q} = \overline{r}, \ldots, R_{d,1,q} = \overline{r}) \equiv p_{\partial,n}(R_{d,k,q} = \overline{r})$ for k to be in $[1, |d|]$.

Although we do not believe the previous assumption to be true in practice because the contexts for making local relevance decisions overlap, this assumption, together with the chain rule, simplifies the joint probability $p_{\partial,n}((R_{d,|d|,q} = \overline{r}), \ldots, (R_{d,1,q} = \overline{r}))$ in Eq. (4) into the sum of the logarithm of the probability of

its individual event. This is as follows.

$$p_{\nabla}(R_{d,q} = r) \propto -\sum_{k=1}^{|d|} \log p_{\partial,n}(R_{d,k,q} = \overline{r}) \tag{5}$$

For occurrences of document terms that are not query terms, we assume that the outcomes of the local relevance decisions for these occurrences are not locally relevant. Using the common string notation that denotes $d[k]$ as the term at the kth location in document d, Wu et al. [2007] called this next assumption query-centric.

Assumption (*Query-Centric*). For any query q and any relevant document d, the relevant information for q is located only in the contexts $c(d, k, n)$ for $k \in [1, |d|]$, where $d[k] \in q$. (i.e., the relevant information is located around query terms).

The preceding assumption is similar to that assumed by the binary independence model [Robertson and Spärck Jones 1976] where nonquery terms in the document are assumed not relevant. The query-centric assumption was corroborated using various TREC ad hoc retrieval test collections [Wu et al. 2007], so it is not validated here.

The query-centric assumption implies that $p_{\partial,n}(R_{d,k,q} = \overline{r}) = 1$ when $d[k]$ is not a query term. This means that $\log p_{\partial,n}(R_{d,k,q} = \overline{r}) = 0$, so Eq. (5) can be simplified by ignoring all locations where query terms do not occur. Using the query-centric assumption and the notation that $Loc(t, d)$ is the set of document locations, given that term $t$occurred in document d (i.e., $t \in d$), Eq. (5) is simplified as

$$p_{\nabla}(R_{d,q} = r) \propto -\sum_{t \in (V(q) \cap V(d))} \sum_{k \in Loc(t,d)} \log p_{\partial,n}(R_{d,k,q} = \overline{r}). \tag{6}$$

2.4 TF-IDF Correspondence

Our nonrelevance decision model in Section 2.3 can be shown to correspond to the TF-IDF term weights as follows. We shrink the context size to unity (i.e., set $n = 0$) based on the following assumption.

Assumption (*Minimal Context*). For any query, the local relevance at a location k in a document d is determined only by the single term $d[k]$.

This assumption is not realistic because the local relevance at location k in document d is not decided by the context, but by the term $d[k]$. From another perspective, such an unrealistic assumption may explain the performance limitations of TF-IDF term weights. Another assumption is that the evaluator makes the same relevance decisions at different locations if the corresponding contexts are the same.

Assumption (*Location-Invariant Decision*). If $c(d, j, n) = c(e, k, n)$, then $\partial_{d,j}(c(d, j, n), q) = \partial_{e,k}(c(e, k, n), q)$ for any query q.

This assumption is used by the document context model [Wu et al. 2007] and was not considered unrealistic. Including the previous two assumptions

implies that the probabilities of local nonrelevance for the same query are the same for different locations, provided that the same term t occurs at these locations. Mathematically, the previous two assumptions imply that if $d[j] = e[k] = t$, then $p_{\partial,0}(R_{d,j,q} = \overline{r}) = p_{\partial,0}(R_{e,k,q} = \overline{r})$. Consequently, we are no longer concerned with the locations of local nonrelevance, but with the presence of query terms in the document. For presentation clarity, we simplify our notation to reflect this as follows.

When the context size is unity (i.e., $n = 0$), the probability of local nonrelevance is

$$p_{\partial,0}(R_{d,k,q} = \overline{r}) = p_{\partial,0}(\partial_{d,k}(c(d,k,0),q) = \overline{r}) = p_{\partial,0}(\partial_{d,k}(t,q) = \overline{r}),$$

where $c(d,k,0) = d[k] = t$. For presentation clarity, we simplify our notation of the previous probability as

$$p_{\partial,0}(\overline{r}|t \in d,q) \equiv p_{\partial,0}(\partial_{d,k}(t,q) = \overline{r}),$$

where the term t occurred in d. The new notation only retains the input and output of the local relevance decision, $\partial_{d,k}(.)$, because it is only based on the term t occurring in d and the query q after the context size is reduced to unity (i.e., $n = 0$). The new notation hides the random variable $R_{d,k,q}$ because d and q have already appeared in the condition part of the probability $p_{\partial,0}(\overline{r}|t \in d,q)$. It also hides the location k because we are no longer concerned with the specific location k of the local nonrelevance, but only with the presence of t in d. The new notation hides the local relevance decision, since this decision is neither directly dependent on the document nor on the location because of the minimal context assumption. Note that the probability by using the new notation is not marginal because it is the probability of local nonrelevance at certain hidden location k where t occurred in d. The location-invariant decision assumption implies that if a term t has an occurrence frequency $f(t,d)$, then there will be an $f(t,d)$ number of times that the same probability $p_{\partial,0}(\overline{r}|t \in d,q)$ appears in Eq. (6). Using this simpler notation, we can rewrite equation (6) as

$$p_{\nabla}(R_{d,q} = r) \propto -\sum_{t \in (V(q) \cap V(d))} f(t,d) \log p_{\partial,0}(\overline{r}|t \in d,q), \tag{7}$$

where $p_{\partial,0}(\overline{r}|t \in d,q)$ is always defined, since t is in $V(q) \cap V(d)$. If Eq. (7) is interpreted as the TF-IDF term weight, then $f(t,d)$ will be the term-frequency factor. The remaining term $-\log p_{\partial,0}(\overline{r}|t \in d,q)$ is called the query-dependent IDF (QIDF).

$$QIDF(t,q) \equiv -\log p_{\partial,0}(\overline{r}|t \in d,q) \tag{8}$$

The following assumption makes the QIDF independent of the query.

Assumption (*Query-Independent Nonrelevance Probability* (*QINRP*)). The conditional probability of nonrelevance, given seeing a term t, is the same for all queries (i.e., $p_{\partial,0}(\overline{r}|t \in d) = p_{\partial,0}(\overline{r}|t \in d,q) = p_{\partial,0}(\overline{r}|t \in d,q')$ for all possible query pairs q and q'.

Note that $p_{\partial,0}(\overline{r}|t \in d)$ is not a marginal probability. Section 4.2 examines the validity of the previous assumption and assesses its impact on retrieval effectiveness.

Assuming that the QINRP assumption is valid, we simplify Eq. (7) to

$$p_{\nabla}(R_{d,q} = r) \propto - \sum_{t \in (V(q) \cap V(d))} f(t, d) \log p_{\partial,0}(\overline{r} | t \in d). \tag{9}$$

For Eq. (9) to correspond to TF-IDF term weights, the remaining quantity (given that t is in the document) after taking $f(t, d)$ away should be the IDF, that is,

$$- \log p_{\partial,0}(\overline{r} | t \in d) = IDF(t). \tag{10}$$

We do not have to consider the case when t is not in d because: (1) $f(t, d)$ is 0, and (2) t must have appeared in d, according to Eq. (9). Section 3.3 derives the previous equation, and therefore establishes Eq. (10). Appendix A has details about the derivation of the term-frequency factor in the literature.

3. INVERSE DOCUMENT-FREQUENCY CORRESPONDENCE

This section shows that the quantity $- \log p_{\partial,0}(\overline{r} | t \in d)$ in Eq. (9) can be approximated by the inverse document-frequency (IDF) [Spärck Jones 1972] as

$$IDF(t) \equiv \log \frac{card(D)}{df(t)},$$

where $df(t)$ is the document frequency of the term t. This approximation simplifies our ranking formula to the TF-IDF term weights.

3.1 Basic Random Match Model

Our approach in this section regards $- \log p_{\partial,0}(\overline{r} | t \in d)$ as a measure of the nonspecificity of term usage of t found in the collection D. Nonspecificity refers to the number of alternatives that one needs to select. Usage refers to the meaning of the term t and the use of t in the context. If the term t occurs at two different document locations with different meanings, then the two usages of t are different. However, the term t at different document locations can have the same meaning, but its usages are still different because the way terms are used can affect the relevance of the usage. For example, the term "telescope" found in two different locations can refer to the same Hubble telescope, but one usage can be about how to repair it and the other about what it has discovered. Therefore, the number of usages of a term is at least the number of meanings that term has in the collection.

The probability $p_{\partial,0}(\overline{r} | t \in d)$ is assigned by our basic random match model of term usages. This model specifies that matching the usage of the query term and the matched document term is done in a random manner, similar to drawing a colored ball from an urn [Feller 1968] at random. In general, more than one usage can be locally relevant to the query, but we make the following assumption to simplify our modeling.

Assumption (*Single Locally Relevant Usage*). A term t has one locally relevant usage for any query out of a set of possible usages of t.

Although this simplifying assumption is not likely to be realistic, it simplifies our basic random match model so that there is only a single parameter to

estimate. If the total number of usages of term t is $m(t)$, then our basic random match model specifies the probability of nonrelevance, given t, as

$$p_{\partial,0}(\overline{r}|t \in d) \equiv \frac{m(t)-1}{m(t)}. \tag{11}$$

Our basic random match model is similar to and inspired by, but not the same as, the probabilistic models based on divergence from randomness [Amati and van Rijsbergen 2002]. To estimate $m(t)$, we estimate the arrival rate $\Lambda(t)$, which is discussed next.

3.2 New-Usage Arrival-Rate Estimation

Consider a hypothetical human evaluator looking up the contexts of our query term t, as in Figure 1, and deciding the relevance of each context to this query. The middle term t in the context is a query term according to the query-centric assumption because contexts of nonquery terms are assumed not locally relevant. The evaluator scans through the set of contexts and collects a set $B(t)$ of unique usages of t from the contexts. Hence $card(B(t)) = m(t)$. A new usage of t is collected if it is different from the set of usages found in $B(t)$ so far. For simplicity, we assume the following.

Assumption (*Poisson Distributed New Term-Usage*). The number of arrivals of new usages of any term in a unit-time interval follows a Poisson distribution.

It follows from the previous assumption that the arrival rate $\Lambda(t)$ of new usages of a term t is a constant. Note that different terms have different arrival rates of new usages.

The conventional estimation of $\Lambda(t)$ counts the number of arrivals of unique usages divided for t by the number of intervals. This estimation is known to be a maximum-likelihood estimator. However, the number of unique usages of a term is not the same as the total number of occurrences of this term, since occurrences of the same term with the same usage are counted only once. Therefore, someone is needed to collect the set of unique usages of a term in the collection, and this collection process is labor intensive and error prone. In addition, the manual identification of similar contexts representing similar term usages can be subjective.

To estimate $\Lambda(t)$ automatically, we regard each document as a constant unit-time interval (which suggests that document lengths should be normalized). If a term is absent in the document, then there will be no new term-usage arrivals in the document. Therefore, we estimate $\Lambda(t)$ by equating the probability that no new term-usages arrived in the document, according to the Poisson distribution with the proportion of documents that do not contain term t as

$$p_{Poisson(\Lambda(t))}(\eta(t)=0) = e^{-\Lambda(t)} = \frac{card(D) - df(t)}{card(D)},$$

where $\eta(t)$ is the number of new term-usages of t, and $p_{Poisson(\Lambda(t))}(.)$ is the probability based on the Poisson model of new-usage arrival. After some algebraic

manipulation, we have an estimate of $\Lambda(t)$.

$$\Lambda(t) = \ln \frac{card(D)}{card(D) - df(t)} \tag{12}$$

We call the previous equation the *zero occurrence* estimate of $\Lambda(t)$. This estimate of $\Lambda(t)$ has a number of problems. First, $\Lambda(t)$ may be a biased estimate. Second, as $df(t)$ approaches $card(D)$, $\Lambda(t)$ tends to infinity. This is because small relative-frequency counts are not reliable estimates of probabilities. Having indicated the problems with this estimate of $\Lambda(t)$, we are not aware of any theoretical alternative to estimate $\Lambda(t)$ without manually identifying the specific usage of each term occurrence. Therefore, we use this estimate of $\Lambda(t)$ assuming that $df(t)$ is not close to $card(D)$ in order to avoid singularities.

3.3 Expectation Approach

Let $E(.)$ be the expectation operator and $\eta(t)$ be the number of unique usages of term t. The expectation approach uses the conditional expected number $E(\eta(t)|\eta(t) > 0)$ of unique usages of term t in document d, given that t occurred in d, as an estimate of the number $m(t)$ of colored balls in an urn in our basic random match model. The conditional expectation is used because the probability $p_{\vartheta,0}(\overline{r}|t \in d)$ in Eq. (11) is a conditional probability where t is present in d. According to the Poisson distributed new term-usage assumption, the number of unique usages follows a Poisson distribution, so the conditional expectation $E(\eta(t)|\eta(t) > 0)$ is calculated as

$$E(\eta(t)|\eta(t) > 0) = \frac{\Lambda(t)}{1 - e^{-\Lambda(t)}},$$

by averaging all possible numbers of new term-usage arrivals in the entire population. Although the number of new term-usage arrivals is bounded by the number of term occurrences in the given document in practice, this bound is not used because the calculated expected number of new-usage arrivals is for the population, and not for a particular document. This treatment is consistent with our minimal context assumption, where $p_{\vartheta,0}(\overline{r}|t \in d)$ depends only on the term and its presence in the document, but not on the particular document d in which t occurred.

Using the previous calculation of $E(\eta(t)|\eta(t) > 0)$, the usages of a term are considered as colored balls drawn from an urn in our basic random match model. Such an urn has $E(\eta(t)|\eta(t) > 0)$ unique usages, where one of the unique usages is assumed the desired usage according to the single locally relevant usage assumption. If the usage of the term in the document is the desired usage matching the usage of t, then the document will be locally relevant to the query. This single local relevance occurrence becomes the document-wide relevance according to the TREC ad hoc evaluation policy [Harman 2004; Clarke et al. 2005]. Likewise, if the usage of the term t is not the usage of the matched query term, then the document location, where the query term occurred in the document, will be locally not relevant to the query. Assuming that each usage of the term t has equal probability of occurrence and using the zero occurrence estimate of $\Lambda(t)$ in Eq. (12), the probability of local nonrelevance for a document

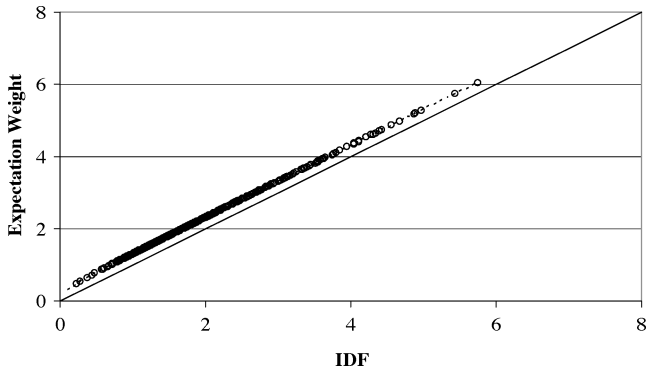

Fig. 2. Relationship between IDF (dotted line) and the expectation weight (solid line). Each circle is the IDF and corresponding expectation weight of a query term in the 200 TREC title queries (see Section 4.1 for details.)

location where the query term t occurred is assigned.

$$p_{\vartheta,0}(\overline{r}|t \in d) \equiv \frac{E(\eta(t)|\eta(t) > 0) - 1}{E(\eta(t)|\eta(t) > 0)} = 1 - \frac{df(t)}{card(D) \ln\left[\frac{card(D)}{card(D) - df(t)}\right]}$$

Using the aforesaid result, we define the expectation weight $W_E(.)$ as a replacement of the IDF for document ranking.

$$W_E(t) \equiv -\log\left[1 - \frac{df(t)}{card(D) \ln\left[\frac{card(D)}{card(D) - df(t)}\right]}\right] \tag{13}$$

In Figure 2, the dotted curve shows the expectation weight, given a specific IDF value. This curve shows the deviation of the IDF value from the expectation weight, since herein the IDF value is supposed to be approximating the expectation weight. In Figure 2, a dotted straight line is drawn to serve as a reference for highlighting the deviation of IDF from the expectation weight (i.e., the solid line). Notice that the IDF value begins to differ from the expectation weight when the former rises above 0.3 (using a logarithm of base 10). This can be explained by deriving the IDF based on a Taylor series expansion of the expectation weight, and will be discussed later.

The circles in Figure 2 represent the IDF values and their corresponding expectation weights of query terms found in the set of 200 TREC title queries in TREC-6, TREC-7, and TREC-2005 ad hoc test collections. Notice that the spread of expectation weights and corresponding IDF values of these query terms are from 0.3 to above 5.0 so that most expectation weights are almost the same as their corresponding IDF values. We observe the difference between the expectation weight and corresponding IDF to slowly increase from 0.2 to 0.3 as the IDF value increases.

We carried out an experiment to observe whether there is any impact on retrieval effectiveness using IDF as an approximation to the expectation weight (Eq. (13)). We used the title queries of TREC-2, TREC-6, TREC-7, and

TREC-2005 ad hoc retrieval test collections. The details of these collections can be found in Section 4.1. In this experiment, the term-frequency factor is based on BM11 [Robertson and Walker 1994], which is multiplied by the IDF, or by the expectation weights, to form the term weights for ranking. We have tested the IDF factor used in the BM11 term weight and the IDF here. Since we could not find any performance differences between them, we did not report their results here.

We measured the retrieval effectiveness of ranking based on IDF (IDF columns) and on expectation weights using data from TREC-2, TREC-6, TREC-7, and TREC-2005 ad hoc retrieval tasks. For all test collections used in this experiment, all the performances are almost the same for ranking based on IDF and on the expectation weights, so numerical details are omitted here. The similar performance may be due to the fact that there are equal numbers of good and bad queries to balance out the performance differences. However, we found that the MAPs of individual queries using ranking based on IDF and the corresponding expectation weights are almost the same. This is substantiated by fitting a linear regression line to the data where the correlation is 1.00 (almost perfect regression), the gradient is 0.9999 (which is approximately 1.0), and the regression curve crosses over the y-axis at 0.00003 (which is close to zero).

We suspect that there are at least two reasons why the retrieval effectiveness of individual queries is similar between ranking based on IDF and the corresponding expectation weights. First, there is almost a constant difference between the expectation weights and IDF values. This difference is about 0.3, small compared with those large expectation weights that usually contribute most in document ranking. Second, if this approximation error of the expectation weight by the IDF value affects all the documents, this error has no impact on ranking. Such approximation errors occur when the document frequency of the query term is large. This implies that almost all the retrieved documents have this query term so that the approximation errors have little impact on ranking the retrieved documents.

In our previous experiment, the IDF was found to be a good approximation of the expectation weights in practice. This good approximation can be shown to hold mathematically. More specifically, the expectation weight in Eq. (13) is simplified to IDF using the Taylor series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad for \quad -1 < x \leq 1$$

by taking only the first term in the Taylor series expansion as follows.

$$\begin{aligned} p_{\partial,0}(\overline{r}|t \in d) &= 1 - \frac{df(t)}{card(D)\ln\left[\frac{df(t)}{card(D)-df(t)} + \cdots\right]} \\ &\approx 1 - \frac{df(t)}{card(D)\frac{df(t)}{card(D)-df(t)}} = \frac{df(t)}{card(D)} \end{aligned}$$

Note that the preceding approximation ($\approx$) and equality ($=$) are not distributive over each other and therefore can only be interpreted as related to the previous

derivation. The aforesaid approximation of $p_{\partial,0}(\overline{r}|t \in d)$ becomes the IDF if we take the negative logarithm of this approximation. The approximation is valid for $-1 \leq df(t)/[card(D) - df(t)] \leq 1$. Therefore, we can simplify the condition for the valid approximation to $df(t) \leq card(D)/2$. Although the major error potentially occurs at the singularity when $df(t) = card(D)$, the quantity $p_{\partial,0}(\overline{r}|t \in d, q)$ tends to 1 in this case. Consequently, the minus logarithm of $p_{\partial,0}(\overline{r}|t \in d, q)$ tends to 0, which is the same as the inverse document-frequency (IDF) value for this particular case (i.e., $\log [card(\mathrm{D})/card(D)]$ for $df(t) = card(D)$). In practice, the previous experiment shows the approximation errors (Figure 2) to have little impact on retrieval effectiveness performance, and this previous condition explains why IDF deviates from the expectation weight when the IDF value is larger than 0.3.

3.4 Inverse Collection Term-Frequency Generalization

The inverse collection term-frequency (ICTF) is proposed by Kwok [1995] and is the same as the inverse location frequency [Roelleke and Wang 2006] for language models. ICTF is conceived as an alternative to IDF, where ICTF counts the individual occurrences of terms instead of the presence or absence of terms in a document for IDF. ICTF is defined as

$$ICTF(t) \equiv \log \frac{tf(D)}{tf(t)},$$

where $tf(D)$ is the total occurrence counts of all terms in the document collection D, and $tf(t)$ is the total occurrence counts of t in the collection.

This subsection shows that the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ can be approximated by ICTF in order to show the generality of our retrieval model (described in Section 2). This approximation is derived in the same way as IDF, using the expectation approach in Section 3.3. However, there is an important difference between the derivations of ICTF and IDF because the former estimates the arrival rate of new term-usages on a per-occurrence basis, whereas IDF estimates this arrival rate on a per-document basis. This difference is consistent with that of the original formulations of ICTF and IDF. We show how the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ is approximated by ICTF, and empirically how ICTF and IDF can be related on the basis that the arrival rates of new term-usages follow a Poisson distribution.

According to the Poisson-distributed new term-usage assumption, the simple change of unit-time interval from a document for IDF to a context for ICTF implies that we can derive ICTF by replacing the document frequency $df(t)$ of term t by the total occurrence frequency $tf(t)$ of t, and the total number of documents by the total number $tf(D)$ of term occurrences of collection D. From this perspective, the zero occurrence estimate of the arrival rate $\Lambda_c(t)$ of new term-usage of t on a per-occurrence basis is

$$\Lambda_c(t) = \ln \frac{tf(D)}{tf(D) - tf(t)}$$

by drawing a parallel with the derivation of Eq. (12).

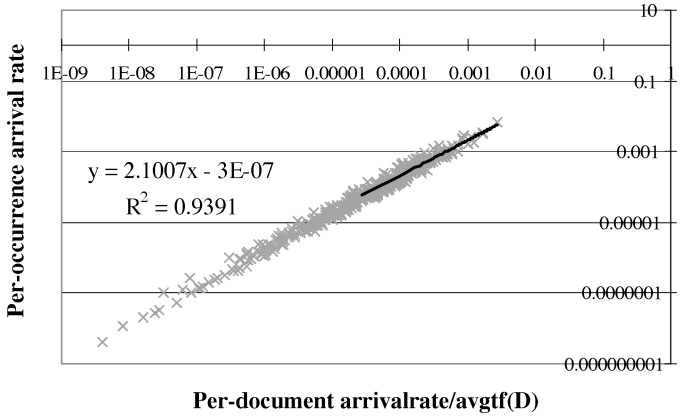

Fig. 3. Scatter diagram of the per-occurrence arrival rate Λ_c of title query terms in the four reference TREC collections (see Section 4.1 for details) against their corresponding per-document arrival rate Λ divided by the average number of terms in a document.

The arrival rate $\Lambda_c(t)$ of new usages of t on a per-occurrence basis is the arrival rate $\Lambda(t)$ of new usages of t on a per-document basis divided by the average number of new term-usages of all terms in the document. It is difficult to know the average number of new term-usages in a document because explicitly counting this number is a labor-intensive and error-prone task. Instead, if the Poisson model of new term-usage arrival is applicable, then the average number of new term-usages will be proportional to the average number $avgtf(D)$ of terms in a document of collection D. Therefore, the arrival rate $\Lambda_c(t)$ of new usages of t on a per-occurrence basis can be specified as

$$\Lambda_c(t) = a\frac{card(D)}{tf(D)}\Lambda(t),$$

where a is the scaling factor that relates the average number of terms in a document to the average number of new-term usages in the document. Note that a is a constant that is independent of a particular term t.

To validate the previous equation that relates the per-occurrence and per-document arrival rates, we plotted the per-document arrival rate divided by the average number of terms in a document against the corresponding per-occurrence arrival rate. This is shown in Figure 3. Each data point in Figure 3 is a TREC title query term in the four reference TREC collections (see Section 4.1). The regression line has a correlation of 96.9%, and passes near the origin (i.e., at 3×10^{-7}). Since the gradient of the regression line is 2.1007, a new usage of a term arrives for every two occurrences of that term, on average. This good curve-fitting result supports the Poisson model of new term-usage arrival in Section 3.2.

Based on the zero occurrence estimate of $\Lambda_c(t)$, we derive the context-counting expectation weight $C_E(.)$ as a replacement of ICTF for document

Table II. Comparison Between IDF and ICTF Using the Wilcoxon Matched-Paris Signed-Ranks Test

	P@10		P@30		MAP		R-Precision	
TREC	IDF	ICTF	IDF	ICTF	IDF	ICTF	IDF	ICTF
2	.438	.426	.399	.379	.193	.175 ($p = 0.0002$)*	.267	.250
6	.388	.384	.284	.283	.218	.219 ($p = 0.1667$)	.266	.261
7	.414	.418	.300	.291	.191	.176 ($p = 0.0698$)	.236	.222
2005	.358	.336	.312	.309	.175	.169 ($p = 0.0414$)	.239	.234

(*) – indicates that the difference in MAP between IDF and ICTF is statistically significant using the Wilcoxon matched-pairs signed-ranks test with 99.9% C. I.

ranking as

$$C_E(t) \equiv \log\left[1 - \frac{tf(t)}{tf(D)\ln\left(\frac{tf(D)}{tf(D)-f(t)}\right)}\right] = -\log p_{\partial,0}(\overline{r}|t \in d)$$

by drawing a parallel with Eq. (13). Taking the first term of the Taylor series expansion of the natural logarithm in the previous equation derives ICTF.

$$-\log p_{\partial,0}(\overline{r}|t \in d) \approx -\log\left[1 - \frac{tf(t)}{tf(D)\frac{tf(t)}{tf(D)-tf(t)}}\right] = \log\frac{tf(D)}{tf(t)}$$

The context-counting expectation weights have the same functional form as the expectation weights. Therefore, the context-counting weights are almost the same as ICTF values, similar to the expectation weights being almost the same as IDF values (see Figure 2). The context-counting weights deviate from ICTF values when $tf(t)$ is larger than half of $tf(D)$ because of the approximation using the Taylor series expansion. For most terms, $tf(t)$ is smaller than half of $tf(D)$ according to Zipf law.

Table II shows the retrieval effectiveness using IDF and ICTF for the four reference TREC collections (see Section 4.1). Using a confidence interval (C.I.) of 99.9%, only TREC-2 shows a statistically significant MAP difference between using IDF and ICTF. However, the MAP differences between using IDF and ICTF are within two percentage points for all four reference TREC collections, so we do not consider their MAPs to differ significantly from each other.

We further test the Poisson distributed new term-usage assumption as follows. If this assumption is correct, the new-usage arrival rate will be scaled from a unit-time interval corresponding to an occurrence to a time interval of a document. Instead of using a single scaling factor a to relate the arrival rate on a per-occurrence basis and the arrival rate on a per-document basis, two scaling factors are needed to relate ICTF and IDF. One factor, b, specifies that $tf(D) = b \times card(D)$, and the other factor, c, specifies that $tf(t) = c \times df(t)$. Two scaling factors are needed because: (1) query terms should not be any terms in the collection, since they exclude stop-words; and (2) one scaling factor is for a single term and the other is for a set of terms. If the factors b and c are

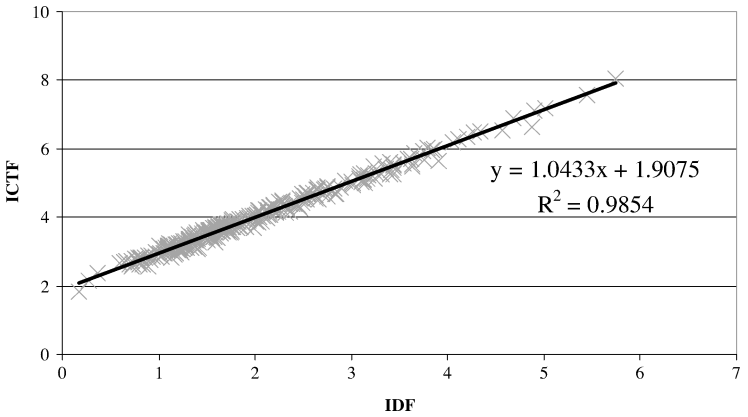

Fig. 4. Scatter diagram of the IDF values and corresponding ICTF values of title query terms in the four reference TREC collections (see Section 4.1 for details).

independent of term t, then ICTF and IDF will be related as

$$ICTF(t) = \log \frac{tf(D)}{tf(t)} = \log \left[\frac{card(D)}{df(t)} \right] + \log \frac{b}{c} = IDF + \log g,$$

where g is a constant combining the two factors b and c. The quantity $\log g$ is a constant independent of query terms, provided that the two factors b and c are constants independent of query terms. The previous equation is similar to the functional form of a simplified version of w_4, which was proposed by Robertson and Walker [1997].

To validate the previous equation empirically, Figure 4 plots a scatter diagram of the IDF- and corresponding ICTF values of over 500 title query terms in the four reference TREC collections (see Section 4.1 for details). A linear regression curve is fitted and the correlation of the regression is 99.3%. The slope of the curve is 1.0433, which is close to 1.0. This supports both our previous equation and our argument that ICTF and IDF are related linearly, based on the assumption that the number of new term-usages follows a Poisson distribution. It supports that the expectation approach can be generalized to derive both IDF and ICTF, as well as for relating IDF and ICTF to the same quantity, $-\log p(\overline{r}|t \in d)$. It also explains how IDF and ICTF differ in regard to their different unit-time intervals of new term-usage arrivals, in the same way as they were originally differently conceived.

3.5 Clustering Approach

The expectation approach presented in previous subsections shows that $p_{\partial,0}(\overline{r}|t \in d)$ can be approximated by IDF, after assuming a random match model of picking a nonrelevant usage and that the new usage of a term is generated by a Poisson process. In this subsection, the clustering approach still assumes the validity of using the random match model, but it does not assume that the new usages are generated by a Poisson process. In addition, it assumes that the number of new usages of t is equal to the number of clusters of similar contexts of t. These clusters are found by a novel clustering algorithm that is described first.

Next we present a more general form of the random match model. Details of the experiments concerning the clustering approach are described in Section 4.

3.5.1 *Context Clustering.* Previous research [Lau and Luk 1999] has identified different usage of a term by clustering the contexts where the term occurred. Results of finding different usage of a term are encouraging, as the performance of identifying these different usages is similar to human identification of various usages. This method of finding different term usages is based on the following assumption.

Assumption (*Similar-Context Similar-Usage*). Terms that have similar usage tend to occur in similar document contexts.

This assumption is similar to the clustering hypothesis [van Rijsbergen 1975] because similar contexts have similar usages and some usages are relevant to a query.

While the results in Lau and Luk [1999] are obtained for Chinese data, we believe that the previous assumption is also valid to the same extent in written languages other than Chinese. This is because word-sense disambiguation algorithms (e.g., Gale et al [1992]) also assign similar senses to a term that is in similar contexts. So, we can treat the problem of estimating the number of usages of a term as the problem of estimating the number of clusters of contexts of a term, where each cluster is assumed to correspond to a unique usage of the term, as in the previous assumption. Using the notation that $v(.)$ returns the vector representation of its argument, $|.|_2$ returns the Euclidean distance of its arguments, and $\bullet$ is the dot product of two vectors, the (cosine) similarity between contexts is computed as follows.

$$sim(v(c(d,k,n)), v(c(d',k',n))) \equiv \frac{v(c(d,k,n)) \bullet v(c(d',k',n))}{|v(c(d,k,n))|_2 \times |v(c(d',k',n))|_2}$$

The weight of term t in $v(c(d, k, n))$ is the TF-IDF term weight (i.e., $f(t, d) \times IDF(t)$).

Hierarchical clustering algorithms are not used here because they do not directly produce the number of clusters. Instead, we use a less popular clustering algorithm based on the idea of the minimum spanning tree (MST) [Zahn 1971; van Rijsbergen 1975]. This algorithm finds a forest, instead of a single tree, that connects all the nodes in the graph. In our case, each node is a context and the edge weight between two nodes is the cosine similarity score between the contexts of these two nodes.

Algorithm 1 shows the major steps in finding the number of clusters. First, the similarity score of each pair of nodes is calculated. Second, these similarity scores are sorted from large to small. Iteratively, the two nodes, say a and b, of the current highest similarity score are checked as to whether they belong to any existing trees formed by the algorithm. If both nodes a and b belong to the same tree, then this tree structure will be destroyed if an edge connecting a and b is added to the tree. Hence, the edge connecting a and b is discarded. If either node a or b is connected to some existing tree, then the existing tree will be extended, with a new edge connecting a and b. If there are no trees that

Algorithm. 1 Modified Minimum Spanning Tree Clustering

```
Step 1   Compute the similarity scores of each pair of nodes
         (or contexts)
Step 2   Sort the similarity scores from large to small
Step 3   From the edge (a, b) with the largest similarity score to the
         smallest do
Step 4       if there is a tree that has both node a and node b then
Step 5         goto step 3 {i.e., skip}
Step 6       if there is a tree that has node a or node b then
Step 7         add (a, b) to the tree
Step 8       else add a new tree with a single edge (a, b)
Step 9       if all the nodes in the graph are connected then goto step 10
Step 10   Count the number of trees as the number m of clusters
Step 11   return m.
```

have nodes a or b, then a and b will form a new tree. This iterative process repeats until all nodes are connected. At the end, the algorithm returns the number of trees formed as the number of clusters found using this modified MST algorithm.

This algorithm assumes that each node is connected with at least one other node. Such a constraint may not be the case if some context (i.e., some node) of a term has a unique usage that no other contexts have. Even if this constraint is not valid, this means that the estimate of the number $m(t)$ of usages of t is less accurate. This constraint affects all terms, so errors due to violation of this constraint are compensated for, to some extent. Since there are also other kinds of errors introduced in the estimation (e.g., similarity score used), the impact of this constraint may not be significant. Experiments detailed in Section 4 investigate whether this clustering algorithm can make good estimates of $p_{\partial,0}(\overline{r}|t \in d)$.

3.5.2 *General Random Match Model.* The general random match model is similar to the basic random match model, except that the former does not make the single locally relevant usage assumption (see Section 3.1). Assuming that the similar-context similar-usage assumption is true, one cluster of similar contexts corresponds to one unique usage, and for a given term t, the number of its different usages is the same as the number $m(t)$ of clusters of similar contexts to t.

For estimations, we make two further simplifying assumptions, as follows.

Assumption (*Equal Probability Cluster*). Each cluster of similar contexts (or each usage) is equally likely to occur.

Suppose that only $h(t, q)$ unique usages (or clusters of similar contexts) out of $m(t)$ are relevant to query q. Also, suppose that the equal probability cluster assumption is true. Then, $p_{\partial,0}(\overline{r}|t \in d, q)$ is the number of unique usages not relevant locally to the query q, divided by the number of unique usages of term t

$$p_{\partial,0}(\overline{r}|t \in d, q) \equiv \frac{m(t) - h(t, q)}{m(t)} \tag{14}$$

because each unique usage, or each cluster of similar contexts, has an equal probability of occurrence according to the equal probability cluster assumption. Note that $m(t)$ is independent of the query because it is the number of possible usages. Given that Eq. (14) is constrained by the algebraic form of Eq. (11) for the random match model, the only variable in Eq. (14) that needs to be dependent on the query is the number $h(t, q)$ of relevant usages to query q.

To estimate $p_{\partial,0}(\overline{r}|t \in d)$, we need to change Eq. (14) to be independent of the query q. The variable $m(t)$ in Eq. (14) depends on the term t, and not on q. The only variable left in Eq. (14) is $h(t, q)$ which is dependent on q. Therefore, to make Eq. (14) independent of q, we parameterize $h(t, q)$ by $\alpha(t)$.

Assumption (*Parameterized Number of Relevant Usage*). For any term t, only $\alpha(t)$ number of usages (or $\alpha(t)$ number of clusters of similar contexts) is relevant to any query and $\alpha(t)$ is independent of the query.

The preceding simplifying assumption implies that the query-independent nonrelevance probability (QINRP) assumption is valid, since Eq. (14) becomes independent of the query when $h(t, q)$ is replaced by $\alpha(t)$. Therefore $p_{\partial,0}(\overline{r}|t \in d)$ is estimated as follows.

$$p_{\partial,0}(\overline{r}|t \in d) \equiv \frac{m(t) - \alpha(t)}{m(t)} \tag{15}$$

Note that when $\alpha(t) = 1$, the estimation of$p_{\partial,0}(\overline{r}|t \in d)$ using Eq. (15) is the same as that of the basic random match model (Eq. (11) in Section 3.1).

Intuitively, when a clustering algorithm only forms tight clusters, probably more than one cluster is relevant to the query and the number of clusters not relevant to the query may be scaled up accordingly. The parameter $\alpha(t)$ can be used to scale back the number of relevant clusters to unity so that the tight clustering effect can be compensated for by $\alpha(t)$. To appreciate this scaling effect, we rewrite Eq. (15) as

$$p_{\partial,0}(\overline{r}|t \in d) = \frac{(m(t)/\alpha(t)) - 1}{(m(t)/\alpha(t))},$$

where $m(t)$ is scaled down to $m(t)/\alpha(t)$, and the number of clusters relevant to the query is always normalized to unity. Experimentally, we can estimate $\alpha(t)$ so that the retrieval effectiveness using the aforesaid equation is similar to that using IDF. Due to the page-length limit, the related experiments are not reported in this article.

4. CLUSTERING APPROACH EXPERIMENTS

This section reports on the experiments of the clustering approach to estimate the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ using the general random match model. Several reference TREC ad hoc retrieval data collections are used.

4.1 Set Up

We test our models with four TREC data collections (i.e., TREC-2, TERC-6, TREC-7, TREC-2005). The TREC-7 documents belong to a subset of the TREC-6 documents. Table III shows some statistics about the data collections and topics (queries) used for the data collections. Title (short) queries are used in

Table III. Statistics of the Collections Used in the Experiments

	TREC-2	TREC-6	TREC-7	TREC-2005
Language	English	English	English	English
Topics	101–150	301–350	351–400	50 past hard topics
No. of documents	714,858	556,077	528,155	1,033,461
No. of relevant documents	11,645	4,611	4,674	6,561
Storage (GB)	3.9	3.3	3.0	5.3

the experiments because they have few (i.e., one to four) query terms, similar to the lengths of Web queries. For statistical inference, we also performed various nonparametric (Wilcoxon) statistical significance tests.

Our retrieval system used the BM11 term weight [Robertson and Walker 1994]. No pseudorelevance feedback is used. All terms in the documents and queries are stemmed using the Porter stemming algorithm [Porter 1980]. Stopwords are removed in both documents and queries.

4.2 Query-Independent Nonrelevance Probability Assumption Validation

Section 2.4 makes three assumptions when the context-based ranking formula given in Section 2.3 is simplified to the basic ranking formula (Eq. (9)). One assumption, the location-invariant decision assumption, is implied by the minimal context assumption when the local relevance decision depends only on the context content, so there are only two assumptions left to validate. In this subsection, we validate one remaining assumption called the query-independent nonrelevance probability (QINRP) assumption. It assumes that the nonrelevance conditional probability $p_{\partial,0}(\overline{r}|t \in d, q)$ depends on the term t and not on the query q because IDF is dependent on t and not on q. The significance of this assumption is that it supports the following.

(1) The minimal context assumption is mainly responsible for the performance degradation and modeling inaccuracies.
(2) This assumption allows derivation of Eq. (9) that forms the basis of the TF-IDF term weights.
(3) It gives the parameterized number of relevant usages assumption of the clustering approach (in Section 3.5.2), for the estimation of $p_{\partial,0}(\overline{r}|t \in d)$ using Eq. (15).
(4) Finally, it is the focus of our subsequent experiments on Eqs. (11) and (15), instead of Eq. (14).

To validate the QINRP assumption, we plot IDF against query-dependent IDF, namely, QIDF, which is based on an estimate of $p_{\partial,0}(\overline{r}|t \in d, q)$ according to Eq. (8). This conditional probability is defined over the event space Ω_∂, so its relative-frequency estimate is the number of nonrelevance contexts divided by the total number of contexts of t. The total number of contexts of t is the total occurrence frequency $tf(t)$ of term t in all documents of the collection because one occurrence of t corresponds to one context. The number of nonrelevance contexts is deduced by subtracting $tf(t)$ from the number of relevant contexts of t for query q. To simplify the estimation of the number of relevant contexts, we supply the following simplifying assumption made by Wu et al. [2007].

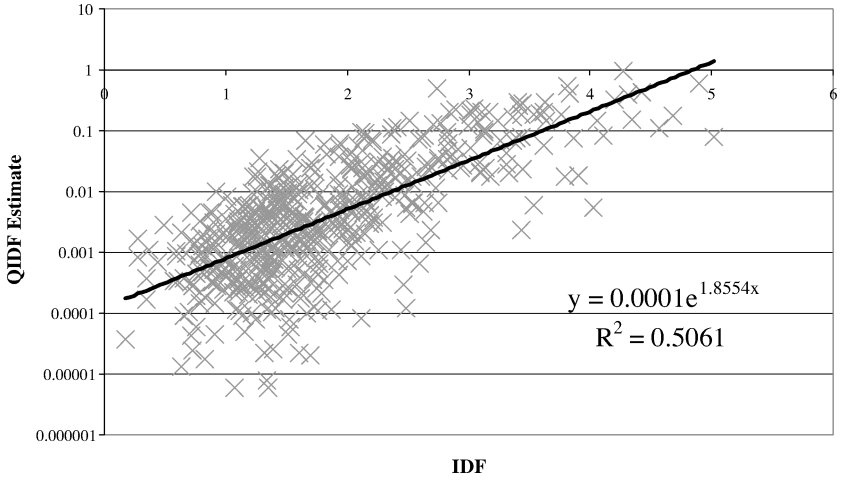

Fig. 5. Scatter diagram of IDF- and corresponding estimated QIDF values of title query terms in the four reference TREC collections.

Assumption (*Context Training*). Given a query q, all contexts of all query terms of q in the relevant documents are relevant.

We make this simplifying assumption even though we know that not every context of a query term in a relevant document is necessarily relevant (see Figure 1, for instance). Using the previous assumption, the conditional probability $p_{\partial,0}(\overline{r}|t \in d, q)$ is estimated by relative-frequency counting as

$$p_{\partial,0}(\overline{r}|t \in d, q) \approx \frac{tf(t) - rtf(t, q)}{tf(t)},$$

where $rtf(t, q)$ is the total occurrence frequency of t in all documents relevant to q. Note that the previous approximation of $p_{\partial,0}(\overline{r}|t \in d, q)$ depends on the query q because $rtf(t, q)$ depends on the query q and this approximation is retrospective because we know which document is relevant to facilitate relative-frequency counting.

Figure 5 plots the IDF and corresponding estimated QIDF of all query terms in the 200 TREC queries of TREC-2, TREC-6, TREC-7, and TREC-2005 ad hoc data collections. It seems that IDF is positively correlated with QIDF. We find that the exponential regression (the solid line in Figure 5) fits the data points with a correlation of 71.1%, which is higher than the correlations of other regression curves that we tried (i.e., linear, logarithmic, and power regression curves). The multiplicative constant in the exponential regression has no impact on ranking because this multiplicative factor is factored out from the basic ranking formula in Eq. (9). However, the exponential function cannot be factored out from Eq. (9), so we cannot replace QIDF by IDF directly. Consequently, we validate the QINRP assumption by examining whether there are any statistically significant differences in retrieval effectiveness using ranking based on QIDF and that based on IDF for the four reference TREC data collections. In this validation, the BM11 term-frequency factor is used.

Table IV. Comparison of Traditional $IDF(t)$ and Context-Dependent IDF ($QIDF(t)$) Performance in Different TREC Data Collections

	P@10		P@30		MAP		R-Precision	
TREC	IDF	QIDF	IDF	QIDF	IDF	QIDF	IDF	QIDF
2	.438	.456	.399	.404	.193	.193 ($p = 0.6328$)	.267	.263
6	.388	.386	.284	.281	.218	.207 ($p = 0.9826$)	.266	.243
7	.414	.412	.300	.296	.191	.183 ($p = 0.5347$)	.236	.222
2005	.358	.342	.312	.300	.175	.175 ($p = 0.7654$)	.239	.237

Table IV shows the retrieval effectiveness of ranking using the basic ranking formula, with an estimate of QIDF and with IDF for the quantity $-\log p_{\partial,0}(\overline{r}|t \in d, q)$ in Eq. (7). The mean average precision (MAP) differences between ranking using QIDF and using IDF are not statistically significant, with a p-value of less than 0.5347 for all four TREC reference collections. This empirically supports the QINRP assumption, at least for the four reference TREC data collections.

For better retrieval effectiveness, we use a more general estimate of QIDF by linearly interpolating the number of these contexts between the maximum number of these contexts (i.e., $rtf(t, q)$) and the minimum number (i.e., $rdf(t, q)$ because each these document must have at least one relevant context according to the TREC ad hoc evaluation policy.) We so interpolate using a mixture parameter α as follows.

$$p_{\partial,0}(\overline{r}|t \in d, q) \approx \alpha \frac{tf(t) - rtf(t, q)}{tf(t)} + (1 - \alpha) \frac{tf(t) - rdf(t, q)}{tf(t)}$$

The best results in terms of retrieval effectiveness using the preceding interpolation formula are obtained when $\alpha = 1.0$ for all four reference TREC data collections. This suggests that the context training assumption is valid for the four reference TREC data collections. Details of the retrieval effectiveness results are not included for presentation clarity.

4.3 Estimating Number of Usages

This section examines whether the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ is better estimated using Eq. (11), which is called the CLU-term weight in this section. Using the modified MST clustering algorithm described in Section 3.5, we obtain the value of $m(t)$ (i.e., number of clusters) for each of the query terms in TREC-6. However, we found there were too many contexts for clustering and the computational resources ran out quickly. To estimate $m(t)$ with less computational resources, we systematically sampled the set of contexts of a term. If the number of contexts is more than 1,000, the systematic sampling ensures that we have a sample of 1,000 contexts. Otherwise, all the contexts are used.

An important parameter when clustering similar contexts is the context size, which in this case is $2n + 1$ terms because there are n terms on each side of the term in the middle of the context. Table V shows the retrieval effectiveness using CLU weight (Eq. (11)) in comparison to IDF for TREC-6 data. The parameter n controlling the context size varies between five and one hundred, but the mean average precision (MAP) values of ranking using the CLU weight differed

Table V. Performance of CLU(t) in TREC-6 with Different Context Sizes Used in the Clustering Algorithm

	n	P@10	P@30	MAP	R-Precision
TREC-6	5	.3460	.2627	.1727	.2086
	15	.3560	.2687	.1836	.2248
	25	.3560	.2660	.1829	.2191
	50	.3520	.2640	.1842	.2210
	100	.3540	.2647	.1835	.2233

Table VI. Comparison of Traditional $IDF(t)$ and Clustering Approach (CLU(t)) Performance in Different TREC Data Collections

	P@10		P@30		MAP		R-Precision	
TREC	IDF	CLU	IDF	CLU	IDF	CLU	IDF	CLU
2	.438	.370	.399	.349	.193	.165 ($p = 0.0048$)	.267	.226
6	.388	.356	.284	.268	.218	.183 ($p = 0.0090$)	.266	.224
7	.414	.362	.300	.252	.191	.167 ($p = 0.0205$)	.236	.213
2005	.358	.288	.312	.272	.175	.153 ($p = 0.0144$)	.239	.211

by no more than one percentage point except for $n = 5$. This suggests that clustering results are insensitive to context size. For efficiency, our subsequent experiments use a context size of 31 (i.e., $n = 15$).

We evaluated the CLU weights using other TREC collections (i.e., TREC-2, TREC-7, and TREC-2005). The retrieval effectiveness of the CLU weights is shown in Table VI. Compared with IDF, the MAPs of the system using CLU-term weights are lower than MAPs of the same system using IDF for all reference TREC data collections. At 99.9% confidence level, none of the collections showed significant difference between the MAP of the system using CLU weights and that using IDF. However, at 99% confidence level, TREC-2 and TREC-6 data showed a significant difference. It seems that CLU is inferior compared with IDF for these cases.

5. RELATED WORK

The probabilistic approach to retrieval was first presented by Maron and Kuhns [1960]. The idea of using probability theory in information retrieval (IR) has generated a number of different competing or complementary probabilistic models [Damerau 1965; Fuhr 1989; Harter 1974; Cooper and Maron 1978; Margulis 1992], such as the binary independence model (BIM) by Robertson and Spärck Jones [1976], the logistic regression model by Cooper et al. [1993; 1992], the TF-IDF term weights by Robertson and Walker [1994] based on the 2-Poisson model [Harter 1975a, 1975b, 1974; Bookstein and Swanson 1974; Robertson et al. 1981], the language model by Ponte and Croft [1998], Zhai and Lafferty [2004], and Lavrenko and Croft [2003, 2001], and more recently the divergence models by Amati and van Rijsbergen [2002]. These models either minimize the (Bayesian) risks (e.g., the BIM and language model [Zhai and Lafferty 2003]), or accept the probabilistic ranking principle (PRP) [Robertson 1977] as the best way to rank documents, maximize the information gain [Amati and van Rijsbergen 2002], or optimize the cross-entropy [Lavrenko and Croft 2003].

Many probabilistic models, as surveyed by Crestani et al. [1998] or as unified by Bodoff and Robertson [2004], as well as retrieval models of other approaches (e.g., the vector space model [Salton et al. 1975]) do not explicitly take into account the term locations in a document, even though term locations have been acknowledged as an important component in determining relevance. For instance, passage retrieval [Kaszkiel et al. 1999; Liu and Croft 2002] implicitly assumes the influence of the query term to be limited within a passage, and local context analysis [Xu and Croft 2000] implicitly assumes that query- and expansion terms are related within some context windows. Language models [Ponte and Croft 1998] use locations to define the location frequencies of term occurrences [Roelleke and Wang 2006], but have not used locations in a more elaborate manner than frequency counting. The question-and-answering (QA) tasks explicitly request the retrieved results to include term locations, but many retrieval models for QA tasks are extensions of existing retrieval models, without explicit consideration of term locations in the model.

We believe that term locations play an important role in determining the relevance of documents to queries. Instead of adding term locations in the retrieval model as a postprocessing module, we develop our probabilistic retrieval model with term locations at the beginning. The local relevance at a certain location is thought to depend on the document context at that location. This is supported by encouraging results in recent studies by Wu et al. [2007, 2006, 2005], as well as by Pickens and MacFarlane [2006] using document-context-based models. By shrinking the context size to unity, we derive the well-known TF-IDF term weights after making some further simplifying assumptions that are similar to the derivations in the language model [Ponte and Croft 1998], the binary independence model [Robertson and Spärck Jones 1976], and the logistic regression model [Cooper et al. 1993, 1992]. From another perspective, these document-context-based models can be thought of as an extension of existing TF-IDF term weights.

Inspired by the divergence model [Amati and van Rijsbergen 2002] that made use of random models, we derived the inverse document-frequency as the information content of the relevance decision (i.e., $-\log p_{\partial,0}(\overline{r}|t \in d)$) when a query term matches a document term. This information content is interpreted as the nonspecificity of term usage. This nonspecificity is derived by assuming that a new usage of a term is generated by a Poisson process, or by counting clusters of similar contexts as clusters of similar usage.

IDF was introduced by Spärck Jones [1972]. It is reasoned on the basis that term occurrences follow a Zipf distribution. A more theoretically motivated term weight w_4 was introduced by Robertson and Spärck Jones [1976] as a generalization of the IDF weights, and w_4 also appears in another context of improving the coordination matching scheme by Yu and Salton [1976]. Since w_4 requires statistics about relevant documents, it is used in retrospective experiments. Croft and Harper [1979] proposed the combination match model (CMM) that relates w_4 with IDF under specific conditions. Later, Robertson and Walker [1997] stated the more general formula (i.e., constant + IDF by Croft and Harper [1979]). This turns out to be similar to the linear mathematical relationship between IDF and ICTF described in Section 3.4. IDF is still

a subject of current research [Joachims 1997; Amati and van Rijsbergen 1998; Hiemstra 1998; Papineni 2001; Aizawa 2003; Roelleke 2003; Lee 2007] where Robertson [2004] and Spärck Jones [2004] responded to recent developments on interpreting IDF. More recent work (e.g., de Vries and Roelleke [2005]) extends the TF-IDF term weights with more elaborate variations. Given the many variations and improvements on the original IDF, this article shows that the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ of our basic ranking formula in Eq. (11) can be approximated by IDF [Spärck Jones 1972] by assuming that the number of new term-usages follows a Poisson distribution.

6. CONCLUSION

This article shows that TF-IDF term weights can be interpreted as making relevance decisions. From this perspective, TF-IDF term weights are the result of simplifying our novel probabilistic retrieval model that simulates human relevance decision-making. This model distinguishes two types of relevance: one common type is the document-wide relevance that applies to the entire document, and the new type is the local relevance that only applies to certain document locations. The model makes local relevance decisions for every location of a document and combines these local relevance decisions into a document-wide relevance decision for the entirety of the document.

The significance of interpreting TF-IDF as making relevance decisions is its potential as a catalyst for different retrieval models and term weights to be interpreted by a unifying perspective: that information retrieval (IR) is about relevance decision-making. Also, our novel probabilistic retrieval model extends TF-IDF term weights to depend on those document locations wherein the query terms occurred. These location-dependent TF-IDF term weights (as in Eq. (6)) have the potential [Wu et al. 2007] to form a basis for developing more elaborate retrieval models for detailed simulation of human relevance decision-making.

Our probabilistic retrieval model ranks documents on the basis of the probability of relevance. Hence, our model complies with the probability ranking principle [Robertson 1977]. When our model is simplified to the basic ranking formula (Eq. (9)), it contains two major factors. The term-frequency factor is the occurrence frequency of the query terms in the document. The remaining quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ is shown IDF if we assume that: (a) a new usage of a term arrives at a constant rate following a Poisson distribution; and (b) the probability of nonrelevance of a given term t is specified by our random match model of term usage. This random match model assumes that: (a) the probability of selecting a particular usage out of a set of possible usages is equally likely; and (b) a term has at most one usage that is relevant to the query. For generality, the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ is also shown ICTF using the same approach to derive IDF, except that the estimate of the new term-usage arrival rate is based on per-occurrence, rather than per-document (as for IDF).

We experimented with another approach that estimates the quantity $-\log p_{\partial,0}(\overline{r}|t \in d)$ for validating our general random match model, without assuming that the new usage of a term arrives at a constant rate following a

Poisson distribution. This approach groups similar contexts into clusters and assumes that similar contexts in a cluster refer to similar usage of the term. We propose a novel modified minimum spanning tree clustering algorithm to find the number of clusters as the number of unique usages of a term. Empirically, we found the retrieval effectiveness of this approach inferior to that using IDF. The problem is that our basic random match model assumed that only one cluster is relevant to a query, but in reality more than one cluster is relevant.

APPENDIX A: TF CORRESPONDENCE

This appendix shows that our term-frequency factor in Eq. (9) can be rendered into different term-frequency factors in the literature [Salton and Buckley 1988; Robertson and Walker 1994] by normlizing the document length. Using the normalized version $\Delta(d)$ of document d, the probability of relevance in Eq. (9) becomes

$$p_{\nabla}(R_{\Delta(d),q} = r) \propto \sum_{t \in (V(q) \cap V(\Delta(d)))} f(t, \Delta(d)) \times IDF(t). \tag{16}$$

The rest of this appendix is organized as follows. Section A.1 describes the basic document length normalization that normalizes the term frequency by the length of document. Section A.2 describes the weighted term-frequency approach that derives the term-frequency factor used by the Okapi system [Robertson 1997]. The weights are determined by the Laplace law of succession when all occurrences of a term in a document are assumed nonrelevant. This approach derives the BM11 term-frequency factor that was used in our previous experiments.

A.1 Proportion Approach

The weighted (Minkowski) p-norm length [Klir and Folger 1988] of d is defined as

$$|d|_p \equiv \sqrt[p]{\sum_w [W(w) \times f(w, d)]^p}$$

with weight $W(w)$ for term w. This weighted p-norm length is related to the weighted generalized mean [Dykchoff and Pedrycz 1984] that is used as the extended Boolean disjunction [Salton et al. 1983]. The vector space model [Salton et al. 1975] uses the weighted Euclidean (i.e., $p = 2$) length and the weight of a term is its IDF. For the unweighted p-norm length, $W(w)$ is set to 1 for all w.

The p-norm length of the normalized document $\Delta(d)$ is denoted by $|\Delta(d)|_p$, which is a constant independent of d. In the literature, $|\Delta(d)|_p$ is the average document length Δ, for $p = 1$. Since $|\Delta(d)|_p$ is a constant, we can deduce the following property of normalized documents.

Property (*Constant Length*). For any two normalized documents, their weighted p-norm lengths are the same, given a particular weighted p-norm.

We define the *p-norm proportion* $g_p(t, d)$ of term t in d as

$$g_p(t, d) \equiv \frac{f(t, d)}{|d|_p}$$

so that we can specify the following assumption.

Assumption (*Constant p-Norm Proportion*). Given a particular weighted p-norm, $g_p(t, \Delta(d)) = g_p(t, d)$ for all terms and for all documents.

Based on the previous assumption, we deduce that

$$f(t, \Delta(d)) = \frac{|\Delta(d)|_p \times f(t, d)}{|d|_p}. \tag{17}$$

Subsituting Eq. (17) into (11), our basic ranking formula becomes

$$p_\nabla(R_{\Delta(d),q} = r) \propto \sum_{t \in (V(q) \cap V(d))} \frac{f(t, d)}{|d|_p} \times IDF(t).$$

It is possible to normalize the query term-frequency as well as using the query length, but we have not pursued this aspect in this article for clarity of presentation.

When $p = 1$, then $|d|_1$ is the number of terms in the document d. The quantity $f(t, d)/|d|_1$ is the relative-frequency estimate of the occurrence probability of term t in document d. When p tends to infinity (i.e., ∞), $|d|_\infty = \max_w\{W(w) \times f(w, d)\}$ [Dykchoff and Pedrycz 1984]. According to the constant-length property, the maximum term-frequency (say, $f_{max} = |\Delta(d)|_\infty$) of all normalized documents is the same (i.e., a constant). When p tends to infinity, the previous ranking formula becomes

$$p_\nabla(R_{\Delta(d),q} = r) \propto \sum_{t \in (V(q) \cap V(\Delta(d)))} \frac{f(t, d)}{\max_w\{W(w) \times f(w, d)\}} IDF(t).$$

When $W(w) = 1$ for all w, the term-frequency factor of the previous equation appears in [Baeza-Yates and Ribeiro-Neto 1999].

We generalize the p-norm proportion approach by linearly interpolating the term frequency of the normalized document and the normalized document length as

$$f(t, \Delta(d)) \equiv |\Delta(d)|_p \times \left[\alpha \times \frac{f(t, d)}{|d|_p} + (1 - \alpha)\right],$$

where α is the mixture parameter. This interpolation captures the intuition that a document without any query terms has small chance of being relevant to the query. This small chance is controlled by α. When $\alpha = 1.0$, the previous equation becomes the normalized term-frequency in Eq. (17) as specified by the p-norm proportion approach. When p tends to infinity and $\alpha = 0.5$, then the previous equation becomes the normalized term-frequency factor by Salton and Buckley [1988].

A.2 Weighted Term-Frequency Approach

Similar to the work by Amati and van Rijsbergen [2002], this approach uses the Laplace law of succession [Feller 1968] to derive the weighted term-frequency (e.g., Huang et al. [2003]) as the term-frequency factor of BM term weights of the Okapi system [Robertson 1997]. This approach derives the BM term weights in a way different from their original conception [Robertson and Walker 1994].

The basic idea is that the term frequency is weighted by a factor $p(f(t,d)|\overline{r})$ that takes into account the probability that all occurrences of term t in document d are locally nonrelevant to a query. This probability is only a weight, and is defined in another event space. Since each occurrence of a term has a weight $p(f(t,d)|\overline{r})$, the term t that occurred $f(t,d)$ times in d has a weighted term-frequency $\omega(t,d)$ of $f(t,d) \times p(f(t,d)|\overline{r})$.

The weight $p(f(t,d)|\overline{r})$ is a probability determined by the Laplace law of succession, as follows. We assume that terms are either locally relevant (r) or locally nonrelevant ($\overline{r}$), corresponding to two outcomes in the Laplace law of succession [Feller 1968]. In this way, $p(f(t,d)|\overline{r})$ is the probability that all the outcomes of $f(t,d)$ occurrences of t are nonrelevant.

$$p(f(t,d)|\overline{r}) \approx \frac{1}{f(t,d)+1}$$

The weighted term-frequency $\omega(t,d)$ of t in d is

$$\omega(t,d) \equiv f(t,d) \times p(f(t,d)|\overline{r}) \approx \frac{f(t,d)}{f(t,d)+1}.$$

Similarly, the weighted, normalized term-frequency $\omega(t, \Delta(d))$ of t in the normalized-length document $\Delta(d)$ is

$$\omega(t, \Delta(d)) \approx \frac{f(t, \Delta(d))}{f(t, \Delta(d))+1}.$$

Assuming that the constant p-norm proportion assumption is true, Eq. (17) is substituted into the previous equation as follows.

$$\omega(t, \Delta(d)) \approx \frac{f(t,d)}{f(t,d)+\frac{|d|_p}{|\Delta(d)|_p}}$$

Replacing $f(t, \Delta(d))$ in Eq. (16) with the previous approximation of $\omega(t, \Delta(d))$ yields the BM11-like [Robertson and Walker 1994] formula as follows.

$$\begin{aligned} p_\nabla(R_{\Delta(d),q} = r) &\propto \sum_{t \in (V(q) \cap V(\Delta(d)))} \omega(t, \Delta(d)) \times IDF(t) \\ &\approx \sum_{t \in (V(q) \cap V(\Delta(d)))} \frac{f(t,d) \times IDF(t)}{f(t,d)+\frac{|d|_p}{|\Delta(d)|_p}} \end{aligned}$$

The previous formula is similar to the BM11 term weight [Robertson and Walker 1994]. First, the original BM11 uses $p = 1$ for measuring document lengths [Robertson and Walker 1994]. Second, the original BM11 has an additive factor, but the highest average precision of the Okapi system is obtained when this additive factor is eliminated (i.e., $k_2 = 0$ in Robertson and Walker

[1994]). Hence, the additive factor is treated as nonexistent in the original BM11 term weight. Third, we do not derive the query term-frequency factor in the original BM11 term weight, for clarity of presentation. Finally, the IDF factor of the original BM11 term weight is w_4 [Robertson and Spärck Jones 1976] for retrospective experiments and becomes IDF1 for predictive experiments, as follows.

$$IDF1(t) \equiv \log\left[\frac{card(D) - df(t) + 0.5}{df(t) - 0.5}\right]$$

We carried out an experiment using the four TREC ad hoc retrieval collections (i.e., same as those in Section 4.1), and found almost no mean average precision differences between ranking using IDF and using IDF1.

The BM25-like term weight [Robertson et al. 1995] is derived by linearly interpolating the original p-norm- and normalized p-norm document lengths with a mixture parameter α, as [Spärck Jones et al. 2000]

$$f(t, \Delta(d)) \equiv \frac{|\Delta(d)|_p \times f(t,d)}{(1-\alpha)|\Delta(d)|_p + \alpha|d|_p}.$$

Substituting the previous equation into $\omega(t, \Delta(d))$, we have

$$\omega(t, \Delta(d)) \approx \frac{f(t,d)}{f(t,d) + (1-\alpha) + \alpha\frac{|d|_p}{|\Delta(d)|_p}}.$$

The BM25-like formula is obtained by substituting the previous equation into our basic ranking formula of Eq. (16).

$$\begin{aligned} p_\nabla(R_{\Delta(d),q} = r) &\propto \sum_{t \in (V(q) \cap V(\Delta(d)))} \omega(t, \Delta(d)) \times IDF(t) \\ &\approx \sum_{t \in (V(q) \cap V(\Delta(d)))} \frac{f(t,d) \times IDF(t)}{f(t,d) + 1 - \alpha + \alpha\frac{|d|_p}{|\Delta(d)|_p}} \end{aligned}$$

The previous formula is similar to the original BM25 term weight [Robertson et al. 1995]. First, the original BM25 has an additive factor, but it was set to zero (i.e., $k_2 = 0$ [Robertson et al. 1995]). Second, the original BM25 term weight includes some multiplicative constants (e.g., $(k_1 + 1)$ and $(k_3 + 1)$ in Robertson et al. [1995]) that do not affect ranking because the additive factor in the original BM25 term has disappeared. Third, the query term-frequency factor of the original BM25 term weight was not derived for clarity of presentation. Finally, the IDF factor of the original BM25 term weight is w_4 [Robertson and Spärck Jones 1976] for retrospective experiments and becomes IDF1 for predictive experiments.

ACKNOWLEDGMENTS

R. W. P. Luk thanks the Center for Intelligent Information Retrieval, University of Massachusetts, for facilitating his development of the basic IR system when he was on leave there. We thank Prof. Robertson and Dr. Amati for improving the article by providing constructive comments.

REFERENCES

AIZAWA, A. 2003. An information-theoretic perspective of TF-IDF measures. *Inf. Process. Manage. 39*, 1, 45–65.

AMATI, G. AND VAN RIJSBERGEN, C. J. 1998. Semantic information retrieval. In *Information Retrieval: Uncertainty and Logics*, C. J. Van Rijsbergen et al., Eds. Kluwer Academic, 189–220.

AMATI, G. AND VAN RIJSBERGEN, C. J. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*, 4, 357–389.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison-Wesley, New York.

BODOFF, D. AND ROBERTSON, S. E. 2004. A new unified probabilistic model. *J. Amer. Soc. Inf. Sci. Technol. 55*, 6, 471–487.

BOOKSTEIN, A. AND SWANSON, D. 1974. Probabilistic models for automatic indexing. *J. Amer. Soc. Inf. Sci. 25*, 312–318.

CALADO, P., RIBEIRO-NETO, B., ZIVIANI, N., MOURA, E., AND SILVA, I. 2003. Local versus global link information in the Web. *ACM Trans. Inf. Syst. 21*, 1, 42–63.

CLARKE, C. L. A. AND SCHOLER, F. 2005. The 2005 terabyte track. In *Proceedings of the 14th Text Retrieval Conference*, Gaithersburg, MD, E. M. Voorhees and L. P. Buckland, Eds. National Institute of Standards and Technology.

CLOUGH, P., SANDERSON, M., AND MÜLLER, H. 2004. The CLEF cross language image retrieval track (ImageCLEF) 2004. In *Proceedings of 3rd International Conference on Image and Video Conference*, Dublin, Ireland, P. Enser et al., Eds. Lecture Notes in Computer Science, vol. 3115. Springer, 243–251.

COOPER, W. S., CHEN, A., AND GEY, F. C. 1993. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In *Proceedings of the 2nd Text Retrieval Conference*, Gaithersburg, MD, D. K. Harman, Ed. National Institute of Standards and Technology, 57–66.

COOPER, W. S., GEY, F. C., AND DABNEY, D. P. 1992. Probabilistic retrieval based on staged logistic regression. In *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, E. Fox et al., Eds. ACM, New York, 198–210.

COOPER, W. S. AND MARON, M. E. 1978. Foundations of probabilistic and utility-theoretic indexing. *J. ACM 25,* 1, 67–80.

CRESTANI, F., LALMAS, M., VAN RIJSBERGEN, C. J., AND CAMPBELL, I. 1998. "Is this document relevant? . . . probably": A survey of probabilistic models in information retrieval. *ACM Comput. Surv. 30,* 4, 528–552.

CROFT, W. B. AND HARPER, D. J. 1979. Using probabilistic models of document retrieval without relevance information. *J. Document. 35,* 285–295.

DAMERAU, F. 1965. An experiment in automatic indexing. *Amer. Document. 16,* 283–289.

DE VRIES, A. P. AND ROELLEKE, T. 2005. Relevance information: A loss of entropy but a gain for IDF? In *Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, G. Marchionini et al., Eds. ACM, New York, 282–289.

DOMBI, J. 1982. A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. *Fuzzy Sets Syst. 8,* 149–163.

DYCKHOFF, H. AND PEDRYCZ, W. 1984. Generalized means as a model of compensative connectives. *Fuzzy Sets Syst. 14,* 143–154.

FELLER, W. 1968. *An Introduction to Probability Theory and Its Applications*, *Vol. 1,* 3rd ed. Wiley, New York.

FRENCH, S. 1986. *Decision Theory:An Introduction to the Mathematics of Rationality*. Ellis Horwood, Chichester, UK.

FUHR, N. 1989. Models for retrieval with probabilistic indexing. *Inf. Process. Manage. 25,* 1, 55–72.

GALE, W. A., CHURCH, K. W., AND YAROWSKY, D. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium Series, Probabilistic Approaches to Natural Language*, Cambridge, MA, 54–60.

HARMAN, D. 2004. Personal communication at NTCIR-4.

HARTER, S. P. 1974. A probabilistic approach to automatic keyword indexing. Ph.D. thesis, Graduate Library, The University of Chicago, Thesis no. T25146.

HARTER, S. P. 1975a. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *J. Amer. Soc. Inf. Sci. 26*, 4, 197–206.

HARTER, S. P. 1975b. A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing. *J. Amer. Soc. Inf. Sci. 26*, 4, 280–289.

HIEMSTRA, D. 1998. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, Heraklion, Crete, Greece, C. Nikolaou and C. Stephanidis, Eds. Springer, London, 569–584.

HUANG, X., PENG, F., SHUURMANS, D., CERCONE, N., AND ROBERTSON, S. E. 2003. Applying machine learning to text segmentation for information retrieval. *Inf. Retr. 6*, 3–4, 333–362.

HUNG, K. Y., LUK, R. W. P., YEUNG, D. S., CHUNG, K. F. L., AND SHU, W. H. 2001. Determination of context window size. *Int. J. Comput. Process. Orient. Lang. 14*, 1, 71–80.

JOACHIMS, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, D. H. Fisher, Ed. Morgan Kaufmann, San Francisco, CA, 143–151.

KASZKIEL, M., ZOBEL, J., AND SACKS-DAVIS, R. 1999. Efficient passage ranking for document databases. *ACM Trans. Inf. Syst. 17*, 4, 406–439.

KLIR, G. J. AND FOLGER, T. A. 1988. *Fuzzy Sets, Uncertainty, and Information*. Prentice-Hall, NJ.

KWOK, K. L. 1995. A network approach to probabilistic information retrieval. *ACM Trans. Inf. Syst. 13*, 3, 324–353.

LAU, K. Y. K. AND LUK, R. W. P. 1999. Word-Sense classification by hierarchical clustering. *J. Chinese Lang. Comput. 9*, 1, 101–121.

LAVRENKO, V. AND CROFT, W. B. 2001. Relevance-Based language model. In *Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, D. H. Kraft et al., Eds. ACM, New York, 120–127.

LAVRENKO, V. AND CROFT, W. B. 2003. Relevance models in information retrieval. In *Language Modeling for Information Retrieval*, W. B. Croft and J. Lafferty, Eds. Kluwer Academic, Chapter 2.

LEE, L. 2007. IDF revisited: A simple new derivation within the Robertson-Spärck Jones probabilistic model. In *Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, C. L. A. Clarke et al., Eds. ACM, New York, 751–752.

LIU, X. AND CROFT, W. B. 2002. Passage retrieval based on language models. In *Proceedings of the 11th ACM Conference on Information and Knowledge Management*, Mclean, VA, C. Nicholas et al., Eds. ACM, New York, 375–382.

LUCASSEN, J. M. AND MERCER, R. L. 1984. An information-theoretic approach to automatic determination of phonemic baseforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA. IEEE, 304–307.

LUHN, H. 1958. The automatic creation of literature abstracts. *IBM J. Res. Devel. 2*, 2, 159–165.

MARGULIS, E. L. 1992. N-Poisson document modelling. In *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, E. Fox et al. Eds. ACM, New York, 177–189.

MARON, M. E. AND KUHNS, J. L. 1960. On relevance, probabilistic indexing and information retrieval. *J. ACM 25*, 3, 216–244.

PAICE, C. D. 1984. Soft evaluation of Boolean search queries in information retrieval systems. *Inf. Technol. Res. Devel. Appl. 3*, 1, 33–41.

PAPINENI, K. 2001. Why inverse document frequency? In *Proceedings of the 2nd Meeting of the North American Chapter of The Association for Computational Linguistics*, Pittsburgh, PA, L. Levin et al., Eds. Association for Computational Linguistics, Morristown, NJ, 25–32.

PICKENS, J. AND MACFARLANE, A. 2006. Term context models for information retrieval. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, Arlington, VA, E. Fox et al., Eds. ACM, New York, 559–566.

PONTE, J. M. AND CROFT, W. B. 1998. A language modeling approach in information retrieval. In *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, W. B. Croft et al., Eds. ACM, New York, 275–281.

PORTER, M. 1980. An algorithm for suffix stripping. *Program 14*, 3, 130–137.
ROBERTSON, S. E. 1977. The probability ranking principle in IR. *J. Document. 33*, 4, 294–304.
ROBERTSON, S. E. 1997. Overview of the Okapi projects. *J. Document. 53*, 1, 3–7.
ROBERTSON, S. E. 2004. Understanding inverse document frequency: On theoretical arguments for IDF. *J. Document. 60*, 5, 503–520.
ROBERTSON, S. E. 2005. On event spaces and probabilistic models in information retrieval. *Inf. Retr. 8*, 2, 319–329.
ROBERTSON, S. E. AND SPÄRCK JONES, K. 1976. Relevance weighting of search terms. *J. Amer. Soc. Inf. Sci. 27*, 3, 129–146.
ROBERTSON, S. E., VAN RIJSBERGEN, C. J., AND PORTER, M. F. 1981. Probabilistic models of indexing and searching. In *Information Retrieval Research*, R. N. Oddy et al., Eds. Butterworths, 35–56.
ROBERTSON, S. E. AND WALKER, S. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, W. B. Croft and C. J. van Rijsbergen, Eds. ACM, New York, 232–241.
ROBERTSON, S. E. AND WALKER, S. 1997. On relevance weights with little relevance information. In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, F. Can et al., Eds. ACM, New York, 16–24.
ROBERTSON, S. E., WALKER, S., AND HANCOCK-BEAULIEU, M. M. 1995. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manage. 31*, 3, 345–360.
ROELLEKE, T. 2003. A frequency-based and a Poisson-based definition of probability of being informative. In *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, C. Clarke et al., Eds. ACM, New York, 227–234.
ROELLEKE, T. AND WANG, J. 2006. A parallel derivation of probabilistic information retrieval models. In *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, S. Dumais, et al., Eds. ACM, New York, 107–114.
SALTON, G. AND BUCKLEY, C. 1988. Term weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24*, 5, 513–523.
SALTON, G., FOX, E. A., AND WU, H. 1983. Extended Boolean information retrieval. *Commun. ACM 26*, 11, 1022–1036.
SALTON, G., WONG, A., AND YANG, C. S. 1975. A vector space model for information retrieval. *J. Amer. Soc. Inf. Sci. 18*, 11, 613–620.
SPÄRCK JONES, K. 1972. Exhaustivity and specificity. *J. Document. 28*, 11–21.
SPÄRCK JONES, K. 2004. IDF term weighting and IR research lessons. *J. Document. 60*, 521–523.
SPÄRCK JONES, K., WALKER, S., AND ROBERTSON, S. E. 2000. A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Inf. Process. Manage. 36*, 6, 809–840.
TROTMAN, A. AND GEVA, S. 2006. Passage retrieval and other XML-retrieval tasks. In *Proceedings of the ACM SIGIR Workshop on XML Element Retrieval Methodology*, Seattle, WA, A. Trotman and S. Geva, Eds., 43–50.
VAN RIJSBERGEN, C. J. 1975. *Information Retrieval*. Butterworths, London.
VECHTOMOVA, O., KARAMUFTUOGLU, M., AND ROBERTSON, S. E. 2006. On document relevance and lexical cohesion between query terms. *Inf. Process. Manage. 24*, 5, 1230–1247.
WALLER, W. G. AND KRAFT, D. H. 1979. A mathematical model of a weighted Boolean retrieval system. *Inf. Process. Manage. 15*, 5, 235–245.
WONG, A. K. C. AND GHAHRAMAN, D. 1975. A statistical analysis of interdependence in character sequences. *Inf. Sci. 8*, 2, 173–188.
WONG, S. K. M., ZIARKO, W., RAGHAVAN, V. V., AND WONG, P. C. N. 1986. On extending the vector space model for Boolean query processing. In *Proceedings of the 9th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Palazzo dei Congressi, Pisa, Italy, F. Rabitti, Ed. ACM, New York, 175–185.
WU, H. C., LUK, R. W. P., WONG, K. F., AND KWOK, K. L. 2005. A retrospective study of probabilistic context-based retrieval. In *Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, G. Marchionini et al., Eds. ACM, New York, 663–664.
WU, H. C., LUK, R. W. P., WONG, K. F., AND KWOK, K. L. 2006. Probabilistic document-context based relevance feedback with limited relevance judgment. In *Proceedings of the 15th ACM*

Conference on Information and Knowledge Management, Arlington, VA, P. S. Yu et al., Eds. ACM, New York, 854–855.

WU, H. C., LUK, R. W. P., WONG, K. F., AND KWOK, K. L. 2007. A retrospective study of a hybrid document-context based retrieval model. *Inf. Process. Manage. 43*, 5, 1308–1331.

XU, J. AND CROFT, W. B. 2000. Improving the effectiveness of information retrieval using local context analysis. *ACM Trans. Inf. Syst, 18*, 1, 79–112.

YAO, Y. Y. AND WONG, S. K. M. 1991. Preference structure, inference and set-oriented retrieval. In *Proceedings of the 14th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL, E. Fox, Ed. ACM, New York, 211–218.

YU, C. T. AND SALTON, G. 1976. Precision weighting—An effective automatic indexing method. *J. ACM 23*, 1, 76–88.

ZAHN, C. T. 1971. Graph-Theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput. 20*, 1, 68–86.

ZHAI, C. X. AND LAFFERTY, J. 2003. A risk minimization framework for information retrieval. In *Proceedings of the ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*.

ZHAI, C. X. AND LAFFERTY, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst. 22*, 2, 179–214.

Received August 2004; revised July 2007; accepted September 2007