Information Processing and Management 49 (2013) 558–575

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Probabilistic co-relevance for query-sensitive similarity measurement in information retrieval

Seung-Hoon Na

Natural Language Processing Research Term, Electronics and Telecommunications Research Institute, South Korea

ARTICLE INFO

Article history:
Received 23 February 2012
Received in revised form 28 September 2012
Accepted 16 October 2012
Available online 24 November 2012

Keywords:
Probabilistic co-relevance
Query-sensitive similarity
Inter-document similarity
Cluster hypothesis
Cluster-based retrieval

ABSTRACT

Interdocument similarities are the fundamental information source required in cluster-based retrieval, which is an advanced retrieval approach that significantly improves performance during information retrieval (IR). An effective similarity metric is *query-sensitive similarity*, which was introduced by Tombros and Rijsbergen as method to more directly satisfy the cluster hypothesis that forms the basis of cluster-based retrieval. Although this method is reported to be effective, existing applications of query-specific similarity are still limited to vector space models wherein there is no connection to probabilistic approaches. We suggest a probabilistic framework that defines query-sensitive similarity based on *probabilistic co-relevance*, where the similarity between two documents is proportional to the probability that they are both *co-relevant* to a specific given query. We further simplify the proposed co-relevance-based similarity by decomposing it into two separate relevance models. We then formulate all the requisite components for the proposed similarity metric in terms of scoring functions used by language modeling methods. Experimental results obtained using standard TREC test collections consistently showed that the proposed query-sensitive similarity measure performs better than term-based similarity and existing query-sensitive similarity in the context of Voorhees' nearest neighbor test (NNT).

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Interdocument similarity is a major factor affecting the enhancement of cluster-based retrieval because it is one of the most crucial factors according to the cluster hypothesis. Typically, *term-based similarity* has been widely used as a similarity metric for inter-document similarity where a document is represented as terms and the similarities between documents then straight-forwardly derived by applying matching functions or retrieval models. They include the Dice coefficient and Jaccard's coefficient, cosine similarity in the vector-space model (Salton, Wong, & Yang, 1975), the probabilistic retrieval model (Robertson & Jones, 1976; Robertson & Walker, 1994), and the KL-divergence between query model and document model in language modeling approaches (Hiemstra, 1998; Lafferty & Zhai, 2001; Ponte & Croft, 1998).

However, it is not known whether term-based similarity is the most suitable metric in terms of the cluster hypothesis. This is because the cluster hypothesis declares the expected properties of *similar* or *closely associated* documents, but does not specify details of the type of interdocument similarities. Therefore, we cannot assume that interdocument similarities necessarily take the form of term-based similarities. Instead of using existing term-based similarity, a more straight-forward way of better similarity metric is to apply the inverse of the cluster hypothesis to produce a similarity metric that might better satisfy the hypothesis.

E-mail address: nash@etri.re.kr

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.10.002

The inspiration for this new method was provided by Tombros and Rijsbergen (2001, 2004), who argued that a form of similarity that better fits the cluster hypothesis should take into account the query context. Based on this argument, they suggested the use of *query-sensitive similarity*, which imparts a query-specific bias on any interdocument similarities found between two documents. This similarity metric makes one pair of documents more similar than others, when both are more similar according to a given query. The similarity obtained is a dynamic quantity that needs to be computed differently for specific goals, in specific retrieval situations, or for each specific query. Evidence for the use of this dynamic similarity metric is found in studies on *query-specific clustering* (Hearst & Pedersen, 1996; Willett, 1985) and *structural re-ranking* on the basis of top-retrieved documents for each query (Kurland & Lee, 2005).

In this pioneering work of Tombros and Rijsbergen (2001, 2004), query-sensitive similarity was only explored in the setting of a vector-space model based on cosine similarity. However, most modern retrieval methods are based on probabilistic frameworks such as the probabilistic retrieval model and language modeling. Therefore, it would be an interesting challenge to develop a query-sensitive similarity method in a probabilistic framework without losing the original insight, and to connect the derived similarity with the cluster hypothesis in a formal manner.

By developing query-specific similarity in a probabilistic context, this study proposes the use of *probabilistic co-relevance* to define similarity more directly and to satisfy the cluster hypothesis. Our main hypothesis is postulated in the co-relevance principle for similarity metric which is stated as follows: *the similarity between two documents should be proportional to the probability in a query context that they are co-relevant to a 'given query' which we call the 'co-relevance probability'*. We consider two different cases; (i) the relevance of a document is independent of the relevance of other documents and (ii) the relevance of a document is dependent on others. We then integrate the resulting estimations from both cases and decompose the co-relevance probability into two *relevance probabilities*. We adopt the same assumptions made in previous works (Lafferty & Zhai, 2003; Roelleke & Wang, 2006), where each relevance probability is further simplified into ranking formulae of retrieval model. Finally, the co-relevance probability is easily and tractably estimated on the basis of simple formulae that rely on top-retrieved documents without resorting to the actual relevant documents. Specifically, we applied language modeling methods for estimating co-relevance probability and obtained query-sensitive similarities which are similar to the interpolation style of RM3, a variant of relevance model (Abdul-jaleel et al., 2004; Lavrenko & Croft, 2001), which is a widely used pseudo-relevance feedback method employed by language modeling approaches.

The results of experiments carried out with standard TREC collections consistently show that the proposed query-specific similarity significantly outperforms the state-of-the-art method developed by Tombros and Rijsbergen (2001) in the setting of Voorhees' nearest neighbor test (NNT) (Voorhees, 1985), thus supporting our claim that co-relevance-based similarity is an improvement over existing metrics.

This paper is organized as follows. Section 2 briefly reviews related works. Section 3 presents our proposed query-sensitive similarity method based on probabilistic co-relevance. Section 4 presents experimental results, and Section 5 presents our conclusions and future work.

2. Related work

Cluster-based retrieval is an approach that generates clusters from collections in order to enhance retrieval performance. This method was motivated by the *cluster hypothesis*, which states that "*closely associated documents tend to be relevant to the same requests*" (Jardine & Rijsbergen, 1971; Rijsbergen, 1979). Thus far, numerous studies have been conducted, for instance, initial trials based on hierarchical clustering that employed different types of merging criteria, i.e., single linkage, complete linkage, group average, and Ward's method (Croft, 1980; El-Hamdouchi & Willett, 1986; Griffiths, Robinson, & Willett, 1984; Jardine & Rijsbergen, 1971; Rijsbergen & Croft, 1975; Voorhees, 1985). There are also more recent language modeling approaches based on partitional clustering (Liu & Croft, 2004; Na, Kang, Roh, & Lee, 2007) and document expansion using nearest neighbors as a cluster (Kurland & Lee, 2004; Tao, Wang, Mei, & Zhai, 2006). The early studies on cluster-based retrieval have delivered inconclusive results; however, recent works based on language modeling methods indicate a significant improvement over the baseline retrieval method (Liu & Croft, 2004; Kurland & Lee, 2004).

Initially, cluster-based retrieval studies were focused on using static clusters from an entire collection as the cluster type. In contrast to these studies, Willett (1985) developed *query-specific clusters*, which were clusters formed from top-retrieved documents in a collection rather than whole documents. However, Willett (1985)'s experiments reported only a limited performance for query-specific clusters when compared with static clusters, possibly due to his limitation in search method as pointed out in Tombros, Villa, and Rijsbergen (2002). Hearst and Pedersen (1996) further motivated the use of query-specific clusters by revising the underlying assumption of the cluster hypothesis and stating that the co-relevance event of two documents should not be fixed statically; it should be dependent on a specific query instead. Hearst and Pedersen (1996) examined the use of query-specific clusters in the setting of an enhanced user interface where a user could choose the best relevant cluster from those recommended by a system. They showed that the retrieval effectiveness can be improved substantially with the interface.

Tombros et al. (2002) further investigated query-specific clusters, but with hierarchical clusters based on four widely used merging methods: single linkage, complete linkage, group average, and Ward's method. They concluded that query-specific clusters showed much better potential effectiveness compared to static clusters.

2.1. Query-sensitive similarity

Query-sensitive similarity was originally motivated by query-dependency of co-relevance, the revised assumption of the cluster hypothesis suggested by Hearst and Pedersen (1996) with the purpose of justifying the use of query-specific clusters, which is stated as follows: "... ··· *we do not assume that if two documents D_1 and D_2 are both relevant or nonrelevant for query Q_A, they must both be relevant or nonrelevant for query Q_B* (Hearst & Pedersen, 1996)".

Tombros and Rijsbergen (2001) noted that this assumption concludes that optimal interdocument similarity should be defined in terms of a query-sensitive function. Based on this insight, they proposed the use of query-sensitive similarity where similarity depends explicitly on a specific query, by imparting a query bias on the interdocument similarity. They suggested that this form of query-sensitive similarity better satisfies the cluster hypothesis.

Query-sensitive similarity was also explored in the field of image retrieval (Zhou & Dai, 2006) and query-oriented summarization (Wei, Li, Lu, & He, 2008). However, as mentioned earlier, all previous formulations of query-sensitive similarity were not based on a probabilistic framework because these formulations were limited to a vector space model. Therefore, it remains unclear whether query-sensitive similarity can be extended to other probabilistic models. This limitation has motivated the major theme of our present research where we aim to incorporate the definition of query-specific similarity in a probabilistic framework.

More recently, Fuhr, Lechtenfeld, Stein, and Gollub (2011) also reversed the cluster hypothesis and proposed the Optimal Clustering Framework (OCF), which was aimed at establishing a 'Principle' for document clustering in a similar way to the Probabilistic Ranking Principle (PRP) in the document ranking. They stated the principle that "*Documents relevant to the same queries should occur in the same cluster*". Based on this principle, Fuhr et al. (2011) developed a novel criterion for clustering that consists of co-relevance probabilities for documents with a more general form, which they generated by summing them over a set of queries not over a specific query as is used in our method. Their experiment showed that the proposed criterion based on co-relevance events is superior to existing clustering criterions in terms of correlation with the true clustering criterion. Given their framework, our proposed co-relevance principle is principally equivalent to OCF in terms of the similarity criterion, if a set of queries is restricted to only a single specific query. However, our subject and goal was slightly different from their study. We were interested in developing a query-sensitive criterion such that a similarity metric is redefined over a *given* specific query, much like Tombros and Rijsbergen (2001)'s work. In contrast, OCF focuses on a *query-based* criterion broadly targeted to a *set of queries*. Furthermore, we also made a more detailed estimation of the co-relevance probability using language modeling approaches, which were not trained in the Fuhr et al.'s (2011) study despite its potential utility. In addition, it should be noted that it becomes harder to *estimate* the co-relevance probability when we focus on the query-sensitive method, which is smaller than the case with a set of queries, because we usually have insufficient training data for a single query. The combination with term-based similarity, which was used in our method, is one way of handling the addressed estimation problem. Some new directions might possibly produce an improved similarity metric by combining query-based and query-sensitive criterions, i.e., using the interdocument similarity obtained with a query-based criterion as a backbone term-based similarity for a query-sensitive criterion.

3. Utilizing probabilistic co-relevance for query-sensitive interdocument similarity

This section first presents an illustrating example of query-sensitive similarity. Next, we present a general description of interdocument similarity based on probabilistic co-relevance, and introduce the problem of co-relevance probability estimation. We then present an approximation for co-relevance probability using probabilistic relevance models that are specifically based on language modeling approaches.

3.1. An illustrating example of a query-sensitive similarity metric

In query-sensitive similarity metric, we place more importance on terms which belong to on a query topic, i.e., we bias a query topic. As a result, the similarity of two documents depends on the context, which means that it can be changed dynamically depending on our focused topic. This metric is contrast to *static* similarity, which the importance of a term is not changed but prefixed in a given collection, such as the inverse document frequency.

In this section, we illustrate an example that motivates the need for a query-sensitive metric. As was demonstrated in the context of image retrieval (Zhou & Dai, 2006), we present a simplified text example illustrating the need for query-sensitive similarity. We assume that a document consists of multiple topics, each topic is given as a single term for simplicity, and the similarity between two documents is computed by the number of matching terms common to both.

Now, suppose that examples of queries q_1 and q_2 and documents d_1, d_2, and d_3 are given as follows.

q_1: a
q_2: b
d_1: $a\ b$
d_2: $a\ d$
d_3: $b\ c$

Given query q_1, d_1 and d_2 are relevant, but d_3 is not relevant. Thus, the d_1 and d_2 pair should both be more similar than the d_1 and d_3 pair. However, the similarity of d_1 and d_3 is the same as that of d_1 and d_2 without query-sensitive similarity. This is because every term is assumed to have equal importance in the original similarity metric. Query-sensitive similarity attributes more weight to terms that belong to a query topic. In the above example, the term a belongs to the query topic, so it should be given more importance when computing the similarity between two documents. Given this query bias, d_1 becomes more similar to d_2 than d_3, since both d_1 and d_2 share the query term a, whereas d_1 and d_3 share the term b which is not a query term. As a result, it is likely that both d_1 and d_2 get more similar relevance scores to q_1 when using query-sensitive similarity, rather than using static similarity for cluster-based retrieval.

In the next subsection, we present the proposed co-relevance principle for interdocument similarity.

3.2. Co-relevance principle for interdocument similarity

Our main objective is to define the similarity metric to best fulfill the cluster hypothesis. To define the similarity, we first adopt the following assumption made by Hearst and Pedersen (1996), which states that the co-relevance of two documents is not a static because it differs depending on the information required.

Query dependency of co-relevance (Hearst & Pedersen, 1996): The co-relevance of two documents depends on the query. Therefore, even if both are relevant to query Q_A, this does not necessarily mean that they must both be relevant to query Q_B.

Given the assumption of query dependency of co-relevance, as stated above, what is the best form of interdocument similarity to satisfy the principle of the cluster hypothesis? The optimum similarity would allow a pair of two documents to be assigned with a maximum similarity when they are co-relevant to a query, but zero otherwise. This kind of co-relevance-based optimal similarity might be obtained, provided we know the set of all documents relevant to a query. Unfortunately, such a set of relevant documents is not available in practice, making it very difficult to compute the exact degree of co-relevance for two documents. Without computing the exact quantity of the degree of co-relevance, we must focus on estimating the probability that two documents are co-relevant to a given query, which we call *co-relevance probability*. Assuming that we know the co-relevance probability, the principle of optimal interdocument similarity for a given query is postulated in the following general statement, which we refer to as the *co-relevance principle for interdocument similarity* or shortly as the *similarity metric principle*.

Co-relevance principle for interdocument similarity (similarity metric principle): For a given query, the interdocument similarity between two documents that best satisfies the cluster hypothesis should be proportional to the probability that they are *co-relevant* to the query.

We now formally present the co-relevance principle. Let $\mathcal{C}$ be the set of documents in the collection. N is the total number of documents in $\mathcal{C}$ (i.e., $|\mathcal{C}|$), and $sim(d,d')$ is the similarity between two documents d and d' ($d,d' \in \mathcal{C}$) to be estimated. Suppose that q is a given query and a binary random variable R denotes the relevance event of a document to q, which is allowed to take one of two possible values, r (relevant) and $\bar{r}$ (nonrelevant). In addition, let $CoRel$ be a binary random variable denoting a co-relevance event of two documents to q, which consists of two possible values $corel$ (co-relevant) and $\overline{corel}$ (non-co-relevant). Thus, the co-relevance principle is formally presented as follows.

$$sim(d,d',q) \propto P(corel|d,d',q) \tag{1}$$

This posits a form of query-sensitive similarity.

3.3. Estimation of the co-relevance probability

The estimation of Eq. (1) requires a subset of relevant documents for a given query q, if there is no simplifying approximation. To dispense with the need for using a set of relevant documents, we focus on a simple estimation method for Eq. (1). Using the chain rule, $P(corel|d,d',q)$ is decomposed to:

$$sim(d,d',q) \propto P(r|d,q)P(r|d',d=rel,q) \tag{2}$$

where $P(r|d',d=rel,q)$ is the probability that d' is relevant to query q, given the condition of the event where d is relevant to q.[1] We need to make a further assumption because the estimation of $P(r|d',d=rel,q)$ still requires a subset of relevant documents for a given query q.

As is based by the cluster hypothesis, the relevance of a document is dependent on those of its similar documents.[2] Thus, we consider two different cases to further derive Eq. (2), i.e., (i) d and d' are similar and (ii) d and d' are not similar. First, if d and d' are similar, we need to use the dependence between the relevance events of given two documents d and d' to decompose $P(r|d',d=rel,q)$. Second, if d and d' are *not* similar, because the cluster hypothesis does not necessarily require that they are *both* relevant or irrelevant to the query, the independence assumption can be taken by simplifying Eq. (2) without any conflict with the cluster hypothesis.

[1] Here, we temporarily introduce ($d=rel,q$) as an event to explicitly indicate that d is relevant to query q.

[2] In this step, we refer to the term "similar documents" as "(significantly) closely associated" documents in the cluster hypothesis, assuming that they are very likely to be co-relevant if they are similar.

However, the proposition that two documents d and d' are similar or not is not *pregiven*, which means, we cannot decide which we choose for further derivation, due to the independence or dependence assumption. In our estimation, a compromising method is used by taking *both* approximations derived from two different cases and *integrating* them into the final probability.

To be more specific, we first consider a case where d and d' are similar. In this case, we remove the condition part on q in $P(r|d', d = rel, q)$, so that $P(r|d', d = rel, q)$ is approximated to $P(d' \text{ is rel}|d \textit{ is rel})$, which means, the conditional probability that d' is relevant to a query, given d is relevant to the query.[3] As a result, Eq. (2) becomes:

$$P_{DM}(corel|d, d', q) \approx P(r|d, q)P(d' \text{ is } rel|d \text{ is } rel) \tag{3}$$

In Eq. (3), we further make a simplifying approximation of $P(d' \text{ is } rel|d \text{ is } rel)$ to $P(r|d', Q = d)$, which means, the probability that d' is relevant to "query document" d. Using this approximation, we obtain the following estimate for $P(corel|d, d', q)$ in this case:

$$P_{DM}(corel|d, d', q) \approx P(r|d, q)P(r|d', Q = d) \tag{4}$$

where we temporarily introduce Q as a random variable to explicitly indicate that d is used as a query.

In the second case, were d and d' are not similar, we make an assumption of independent relevance. Given this simplification, Eq. (2) is estimated further as follows.

$$P_{IM}(corel|d, d', q) \approx P(r|d, q)P(r|d', q) \tag{5}$$

Thus, we now have two different estimates, i.e., Eqs. (4) and (5). Without determining which approximation is more suitable for $P(corel|d, d', q)$, we suggest that we geometrically interpolate these estimates $P_{DM}(corel|d, d', q)$ and $P_{IM}(corel|d, d', q)$ for estimation. As a result, we obtain an estimate of $P(corel|d, d', q)$, as follows:

$$P(corel|d, d', q) \approx P_{DM}(corel|d, d', q)^{1-\alpha}P_{IM}(corel|d, d', q)^{\alpha} = P(r|d, q)(P(r|d', Q = d)^{(1-\alpha)}P(r|d', q)^{\alpha}) \tag{6}$$

where α is a controlling parameter that indicates the relative weight of the case for the independence assumption. For convenience, we introduce two notations, i.e., $sim_{TSM}(d, d')$ and $sim_{QSSM}(d, d', q)$, to aggregate the components of Eq. (6) into query-independent and dependent groups, respectively, which are defined as:

$$sim_{TSM}(d, d') = P(r|d', Q = d) \tag{7}$$
$$sim_{QSSM}(d, d', q) = P(r|d, q)P(r|d', q) \tag{8}$$

Using the notations $sim_{TSM}(d, d')$ and $sim_{QSSM}(d, d', q)$, Eq. (6) is now reformulated as:

$$\begin{aligned} sim_{TSM+QSSM}(d, d', q) &\propto P(corel|d, d', q) = P(r|d', Q = d)^{(1-\alpha)}P(r|d, q)^{1-\alpha}\left(P(r|d, q)P(r|d', q)\right)^{\alpha} \\ &= P(r|d, q)^{1-\alpha}sim_{TSM}(d, d')^{1-\alpha}sim_{QSSM}(d, d', q)^{\alpha} \propto sim_{TSM}(d, d')^{1-\alpha}sim_{QSSM}(d, d', q)^{\alpha} \end{aligned} \tag{9}$$

Thus, the final proposed estimation Eq. (9) takes the form of a combined metric with two different similarities, where one is $sim_{TSM}(d, d')$ which is known as *term-based similarity*, while the other is $sim_{QSSM}(d, d', q)$ which is known as (purely) query-sensitive similarity. Unless specified otherwise, *query-sensitive similarity* refers to either the combined metric $sim_{TSM+QSSM}(d, d')$, or $sim_{QSSM}(d, d', q)$. Note that the co-relevance probability is not a symmetric metric, because approximations using Eqs. (3) and (4) are not "derived" forms but being approximated, and $P(r|d', Q = d)P(r|d, q)$ is not the same as $P(r|d, Q = d')P(r|d', q)$ in general.[4]

3.3.1. Estimation of the relevance probability

The co-relevance probability is now decomposed into a relevance probability $P(r|d, q)$ (or $P(r|d', Q = d)$), which is the *probability* that d is *relevant* to q. This simplification allows us to readily manipulate $P(corel|d, d', q)$ because $P(r|d, q)$ is grounded in established IR models such as probabilistic retrieval models and language modeling (Lafferty & Zhai, 2003; Roelleke & Wang, 2006). In other words, this simplified setting allows $P(r|d, q)$ to be derived without a set of relevant documents by exploiting approximated derivations from a probabilistic retrieval model, which is well-known in the IR literature (Lafferty & Zhai, 2003; Roelleke & Wang, 2006).[5]

As shown in Roelleke and Wang (2006), we rewrite the relevance probability $P(r|d, q)$ using Bayes' rule as follows.

$$P(r|d, q) = \frac{P(d, q, r)}{P(d, q)} = \frac{P(d, q, r)}{P(d, q, r) + P(d, q, \bar{r})} \tag{10}$$

[3] $P(d' \text{is rel}|d \textit{ is rel})$ is not dependent on a specific query or a specific information need.

[4] This depends on whether a retrieval model satisfies the symmetric constraint $P(r|d', Q = d)P(r|d, q) = P(r|d, Q = d')P(r|d', q)$. For example, a vector-space model using cosine-similarity will guarantee this symmetric constraint, although it requires a probabilistic interpretation of the relevance model.

[5] There might be some superficial confusion here, because our simplified setting for the relevance of a document is *independent* of other documents when deriving Eq. (5), which was originally adopted in the Probabilistic Ranking Principle (Robertson, 1977). This might appear to be a logical conflict with the cluster hypothesis, which assumes that relevance should be *dependent* on other documents, as shown by Goffman (1968). However, we do not assume *independence* to satisfy the cluster hypothesis, but instead we use the *independence* of relevance as an approximation only when simplifying our similarity metric method.

Further derivations are found in Roelleke and Wang (2006), but we focus on the odd set of the relevant event $O(r|d,q)$, which is a fraction of the probabilities of relevant and nonrelevant events. This is given as follows.

$$O(r|d,q) = \frac{P(d,q,r)}{P(d,q,\bar{r})} \tag{11}$$

$p(r|d,q)$ can be represented by $O(r|d,q)$ using the following relationship.

$$P(r|d,q) = \frac{O(r|d,q)}{O(r|d,q)+1} \tag{12}$$

Depending on the use of query generation (language modeling approaches) or document generation (conventional probabilistic retrieval model), Eqs. (11) and (12) yield different formulae. In a case study, this paper introduces formulae using query generation with language modeling approaches, which are also found in existing studies. Our derivations introduced here are all based on the underlying probabilistic semantics found in Lafferty and Zhai (2003), which are partly based on assumptions found in Roelleke and Wang (2006).

3.3.2. Language modeling approach

In the language modeling approaches, $P(d,q,r)$ and $P(d,q,\bar{r})$ are factored into the following formulae, based on the semantics of Lafferty and Zhai (2003), as follows.

$$\begin{aligned} P(d,q,r) &= P(q|d,r)P(d|r)P(r) \\ P(d,q,\bar{r}) &= P(q|d,\bar{r})P(d|\bar{r})P(\bar{r}) \end{aligned} \tag{13}$$

The query generation processes are explicitly stated when using $P(q|d,r)$ and $P(q|d,\bar{r})$, which are the generative probabilities that q becomes a sample to be generated. Thus, we formulate this using a ranking principle, which is based on the query likelihood used by language modeling approaches. We assume unigram generation for $P(q|d,r)$ and $P(q|d,\bar{r})$, which are further derived as follows.

$$\begin{aligned} P(q|d,r) &= \prod_{w\in q} P(w|d,r)^{c(w,q)} \\ P(q|d,\bar{r}) &= \prod_{w\in q} P(w|d,\bar{r})^{c(w,q)} \end{aligned} \tag{14}$$

where $c(w, q)$ is the term frequency of w in the query q. For brevity, we regard q as both a set and a sequence of words.

We simply refer to $P(w|d,r)$ and $P(w|d,\bar{r})$ as the *relevant document model* and the *nonrelevant document model*, which describe the generative model for relevant documents and nonrelevant documents, respectively.

As suggested by Hiemstra (1998, 2001), Miller, Leek, and Schwartz (1999), and Roelleke and Wang (2006), we first regard the relevance document model $P(w|d,r)$ as a linear mixture of the document language model $P(w|d)$ and the collection model $P(w|\mathcal{C})$, as follows.

$$P(w|d,r) \approx (1-\lambda_d)P(w|d) + \lambda_d P(w|\mathcal{C}) \tag{15}$$

MLE is used to estimate $p(w|d)$, while λ_d is a mixture parameter. Depending on the smoothing method (Zhai & Lafferty, 2001), λ_d might be dependent or independent of d.

The document-specific term in the nonrelevance model $P(w|d,\bar{r})$ is eliminated by assuming the conditional independence of document d and query q, with the nonrelevant event $\bar{r}$, following the simplifying assumption of Lafferty and Zhai (2003). Thus, $P(q|d,\bar{r}) = P(q|\bar{r})$. We aggressively further assume $P(w|d,\bar{r}) = P(w|\bar{r})$, and $P(w|\bar{r})$ is simply approximated by $P(w|\mathcal{C})$. This is because most of the documents in a collection are nonrelevant to a query. This leads us to a final formulation of $P(w|d,\bar{r})$ as follows.

$$P(w|d,\bar{r}) = P(w|\bar{r}) \approx P(w|\mathcal{C}) \tag{16}$$

Note that the derivations of Eqs. (15) and (16) employ the background collection model $P(w|\mathcal{C})$ to handle the relevance and nonrelevance models. This means that $P(w|\mathcal{C})$ plays dual roles in relevance and nonrelevance model; More specifically, $P(w|C)$ is used for "smoothing" in the relevance document model, whereas $P(w|C)$ "approximates" the nonrelevance document model. This duality of the collection model was also exploited by Lavrenko and Croft (2001) and Lavrenko (2010) when estimating the relevance and nonrelevance models.[6]

We now complete the formulation of $O(r|d,q)$. First, we decompose the odd set $O(r|d,q)$ as follows.

[6] In contrast to our method and those of Lavrenko (2010) and Lavrenko and Croft (2001), Roelleke and Wang (2006) only used one part for both roles in the setting of their relevance assumption, i.e., the relevance role of the collection, because they stated that the collection model constituted a relevance document model for the language modeling approach. Roelleke and Wang (2006) did not consider duality, but they had a different objective to that of our study. Previous studies have focused on a ranking-equivalent form of $O(r|d,q)$, which means that $P(q|d,\bar{r})$, the denominator part of the odd set (Lafferty & Zhai, 2003; Roelleke & Wang, 2006), need not be derived further because it is reduced to a document-independent term, i.e., $P(q|\bar{r})$ with no effect on the ranking. In contrast, our work requires the co-relevance probability, which demands a full formulation of $O(r|d,q)$ and $P(q|\bar{r})$ without eliminating any of its components.

$$O(r|d,q) = \frac{P(q|d,r)P(d,r)}{P(q|d,\bar{r})P(d,\bar{r})} = \frac{P(q|d,r)P(r)P(d|r)}{P(q|d,\bar{r})P(\bar{r})P(d|\bar{r})} \tag{17}$$

By substituting Eqs. (14)–(16) for Eq. (17), the final form of $O(r|d,q)$ is given as follows.

$$O(r|d,q) = \frac{P(r)P(d|r)}{P(\bar{r})P(d|\bar{r})} \lambda_d^{|q|} \prod_{w \in q} \left(\frac{(1-\lambda_d)P(w|d)}{\lambda_d P(w|\mathcal{C})} + 1 \right)^{c(w,q)} \tag{18}$$

where $|q|$ is the length of q. Note that $O(r|d,q)$ is rank-equivalent to the retrieval method using Jelinek–Mercer or Dirichlet prior smoothing, as found in language modeling approaches (Zhai & Lafferty, 2001), given the assumption of $P(d|r) = P(d)$ and $P(d|\bar{r}) = P(d)$ introduced by Lafferty and Zhai (2003), or to their further extensions using document priors (Blanco & Barreiro, 2008; Hiemstra, 1998; Losada & Azzopardi, 2008; Smucker & Allan, 2005), given the assumed prior probability settings for $P(d|r)$ or $P(d|\bar{r})$.

3.3.3. Special case: Dirichlet prior smoothing

The form of Eq. (18) contains a smoothing parameter λ_d. This section considered a specific variant of Eq. (18) using Dirichlet prior smoothing, which a very popular smoothing method in language modeling because of its high effectiveness with short keyword queries (Zhai & Lafferty, 2001).

With Dirichlet-prior smoothing, λ_d is defined by $\mu/(\mu + |d|)$ using a smoothing parameter μ. During Dirichlet-prior smoothing, the relevance score of a document d for a given query q is formulated as follows (Zhai & Lafferty, 2001).

$$S_q^{LM}(d) = \sum_{w \in q} c(w,q) \log \left(\frac{(1-\lambda_d)P(w|d)}{\lambda_d P(w|\mathcal{C})} + 1 \right) + |q| \log(\lambda_d) \tag{19}$$

Using $S_q^{LM}(d)$, we readily show that Eq. (18) can be rewritten as follows.[7]

$$O(r|d,q) = \frac{P(r)}{P(\bar{r})} \exp\left(S_q^{LM}(d)\right) \tag{20}$$

This gives the following equivalent form for $P(r|d,q)$ from Eq. (12).

$$P(r|d,q) = \frac{\exp\left(S_q^{LM}(d)\right)}{\exp\left(S_q^{LM}(d)\right) + P(\bar{r})/P(r)} \tag{21}$$

To further simplify the formula, let K be $p(\bar{r})/p(r)$. By substituting Eq. (21) to Eq. (8), we obtain the final form of query-specific similarity as follows:

$$sim_{QSSM}(d,d',q) \propto \left(\frac{\exp\left(S_q^{LM}(d)\right)}{\exp\left(S_q^{LM}(d)\right) + K} \right) \left(\frac{\exp\left(S_q^{LM}(d')\right)}{\exp\left(S_q^{LM}(d')\right) + K} \right) \tag{22}$$

Using the same derivation, we obtain the final form of the term-based similarity as follows.

$$sim_{TSM}(d,d') \propto \left(\frac{\exp\left(S_d^{LM}(d')\right)}{\exp\left(S_d^{LM}(d')\right) + K} \right) \tag{23}$$

As a result, the query-specific similarity given by Eq. (22) is precisely expressed with standard scoring functions in language modeling approaches $S_q^{LM}(d)$. It is simple to show that this query-specific similarity is symmetric, i.e., $sim(d,d') = sim(d',d)$.

As a result, the query-specific similarity and the term-based similarity given by Eqs. (8) and (7) are precisely expressed using standard scoring functions in language modeling approaches $S_q^{LM}(d)$ and $S_d^{LM}(d')$, respectively.

One remaining practical concern is that $P(q|d,r)$ (or $P(q|d,Q = d')$) is significantly smaller if then query length is larger, which possibly means that highly relevant documents have low relevance probabilities. To ameliorate the effect of query length, we use its *length-normalized query* by dividing the term frequencies by the original query length such that the resulting length becomes the unit length, instead of using the original query. A similar kind of length normalization for documents or queries has also used in the language modeling literature (Kurland & Lee, 2005; Lavrenko et al., 2002).

Formally, let the length-normalized query for q be $\tilde{q}$. The term frequency for each word in query $\tilde{q}$ is then defined as follows.

$$c(w,\tilde{q}) = P_{ml}(w|q) = \frac{c(w,q)}{|q|} \tag{24}$$

[7] Here, $P(d|r)$ and $P(d|\bar{r})$ are dropped, as we assume that $P(d,r) = P(d)P(r)$, and $P(d,\bar{r}) = P(d)P(\bar{r})$, by adopting the derivation step introduced by Lafferty and Zhai (2003).

$P_{ml}(w|q)$ is the MLE of the language model for query q. By using $\tilde{q}$ instead of the original query q, Using the length-normalized query, $P(r|d,\tilde{q})$ is reformulated as follows.

$$P(r|d,\tilde{q}) = \frac{\exp\left(S_q^{LM}(d)\right)}{\exp\left(S_q^{LM}(d)\right) + K \cdot \exp(|q|)} \tag{25}$$

which finally leads to the following length-normalized formula for $sim_{QSSM}(d,d',q)$ and $sim_{TSM}(d,d')$:

$$sim_{QSSM}(d,d',q) = \left(\frac{\exp\left(S_q^{LM}(d)\right)}{\exp\left(S_q^{LM}(d)\right) + K \cdot \exp(|q|)}\right)\left(\frac{\exp\left(S_q^{LM}(d')\right)}{\exp\left(S_q^{LM}(d')\right) + K \cdot \exp(|q|)}\right) \tag{26}$$

$$sim_{TSM}(d,d') = \left(\frac{\exp\left(S_d^{LM}(d')\right)}{\exp\left(S_d^{LM}(d')\right) + K \cdot \exp(|d|)}\right) \tag{27}$$

3.3.4. Simple setting for P(r)

In Eq. (18), the remaining issue is how to set $P(r)$. To estimate $P(r)$, this section presents a simplified method using an example case, where $P(r)$ is assumed to be *extremely* small. Based on this assumption, $P(\bar{r})/P(r)$ is very large, so the denominator parts of $P(r|d,q)$ are almost independent of document d, which leads us to obtain the following simplified form of Eq. (6), which leads us to make $P(r|d,q)$ be proportional to $O(r|d,q)$.[8] As a result, we obtain the following simplified form for Eq. (6).

$$P(corel|d,d',q) \propto\propto O(r|d,\tilde{q})O(r|d',\tilde{d})^{1-\alpha}O(r|d',\tilde{q})^{\alpha} \tag{28}$$

This is referred to as the *odd-based co-relevance model.*

Taking a log function of Eq. (28) provides the following variant.

$$P(corel|d,d',q) \propto\propto (1-\alpha)\frac{S_d^{LM}(d')}{|d|} + \alpha\frac{S_q^{LM}(d')}{|q|} + \frac{S_q^{LM}(d)}{|q|} \tag{29}$$

4. Experimentation

This section presents experiments that compare the proposed co-relevance-based similarity with term-based similarity.

4.1. Baseline: Tombros and Rijsbergen's query-specific similarity

The original formula given by Tombros and Rijsbergen (2001) is based on a vector space model, which is simply referred to as **TR similarity** in this paper. The major difference between our similarity method and TR similarity is that we add probabilistic co-relevance to the original framework. This means that the query-specific metric can be applicable to other probabilistic models.

We formally presents the TR similarity by first presenting the general vector space model. Suppose that $W(\cdot,d)$ and $W(\cdot,q)$ are weighting functions defined for document d and q, and that $\|d\|$ and $\|q\|$ are norms of d and q, respectively. The relevance score of document d to query q is then formulated in the vector space model as follows.

$$S_q^{VSM}(d) = \sum_w \frac{W(w,d)}{\|d\|}\frac{W(w,q)}{\|q\|} \tag{30}$$

Based on the original definition of TR similarity (Tombros & Rijsbergen, 2001), if we are given two documents d and d', we first need to artificially define their common document $d \otimes d'$, which we call a *co-representation* of d and d', i.e., a virtual document comprised of the shared terms among d and d'. The TR similarity $sim(d,d',q)$ is then defined using a co-representation as the similarity score between query q and the common document $d \otimes d'$, as follows.

$$sim(d,d',q) = S_q^{VSM}(d \otimes d') \tag{31}$$

By substituting Eq. (30) to Eq. (31), the final generalized form of the TR similarity is given as follows.

$$sim(d,d',q) = \sum_w \frac{W(w,d \otimes d')}{\|d \otimes d'\|}\frac{W(w,q)}{\|q\|} \tag{32}$$

[8] For the approximation, note that we focus on the relative ratio between two similarity values, i.e., $sim(d,d_1,q)/sim(d,d_2,q)$, rather than the absolute values. More specifically, suppose that we are interested in R, the ratio of $S_1/(S_1 + K)$ over $S_2/(S_2 + K)$, i.e., $R = (S_1/S_2) \times (S_2 + K)/(S_1 + K)$ where S_1 and S_2 are score values. When K is large, part of the ratio $(S_2 + K)/(S_1 + K)$ will converge to 1, so the ratio R becomes S_1/S_2. Equivalently, this means that only the nominator parts (i.e., the score parts) will affect the similarities.

The remaining problem is how to define the co-representation, i.e., the term frequency of w for the pseudo document $d \otimes d'$. In the original study (Tombros & Rijsbergen, 2001), the term frequency of a word in a co-representation was defined as the *geometric mean* of the individual term frequencies of the word as follows.

$$c(w, d \otimes d') = \sqrt{c(w,d) \cdot c(w,d')} \tag{33}$$

Another issue that remains is how to specify the weights of the document, the query (i.e., $W(w,d)$ and $W(w,q)$), and their norms (i.e., $\|d\|$ and $\|q\|$). Originally, TR similarity was based on SMART's *ltc* scheme for specifying the weights of the document, the query, and their norms, and the formula is summarized in Buckley, Salton, Allan, and Singhal (1994). Under the ltc scheme, the weights of document d and query q are given as follows.

$$W(w,d) = \begin{cases} \log(c(w,d)) + 1 & c(w,d) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

$$W(w,q) = c(w,q) \log\left(\frac{N+1}{df(w)}\right) \tag{35}$$

where N is the total number of documents in the collection, and $df(w)$ is the document frequency of w. Under the ltc scheme, the norms of the document and the query are given as follows.

$$\begin{aligned} \|d\| &= \sqrt{\sum_w W(w,d)^2} \\ \|q\| &= \sqrt{\sum_w W(w,q)^2} \end{aligned} \tag{36}$$

It is simple to show that applying Eqs. (33)–(36) to Eq. (32) yields exactly the same formula as that described in the original paper.[9]

However, the ltc scheme has weak performance and is therefore considered to be far from a state-of-the-art vector space model. To produce a more convincing experimental result, it would be more useful to use a *pivoted vector space model* based on the *pivoted length normalization method*, which is the best performing vector space model (Singhal, Buckley, & Mitra, 1996). According to Singhal (2001), the pivoted vector space model is based on the following document and query weights.

$$W(w,d) = \begin{cases} 1 + \log(\log(c(w,d)) + 1) & c(w,d) > 0 \\ 0 & otherwise \end{cases} \tag{37}$$

$$\begin{aligned} \|d\| &= (1-s) + s\frac{|d|}{avgl} \\ \|q\| &= 1 \end{aligned} \tag{38}$$

where $W(w,q)$ in the pivoted vector space model is the same as Eq. (35), $|d|$ is defined by $\sum_w c(w,d)$, s is the pivoted parameter, and *avgl* is the average length of documents in the collection.[10] Pivoted vector space model allows the formula for query-specific similarity to be readily derived by applying Eqs. (37) and (38) to the original definition and using the co-representation formula of Eq. (33).

Tombros and Rijsbergen (2004) further suggested a combined form of query-specific similarity defined above as $sim(d,d',q)$, using a term-based similarity metric. In this combination, two variants were suggested, i.e., multiplication and linear interpolation. Their experimental results showed that these variants delivered comparable performances. In this work, we use the linear interpolation, which is denoted as **M3** in the work of Tombros and Rijsbergen (2004).

To complete the formalization, we define the term-based similarity $sim_{TSM}(d,d')$ as the relevance score of document d' for query document d, using the same weighting function found in Eq. (30).

$$sim_{TSM}(d,d') = S_d^{VSM}(d') \tag{39}$$

The combined metric is defined as follows.

$$sim_{TSM+QSSM}(d,d',q) = (1-\beta)sim_{TSM}(d,d') + \beta sim_{QSSM}(d,d',q) \tag{40}$$

where β is an interpolation parameter.

[9] The aim of our derivation is a generalization of the original method, so, although it appears different from that described in original paper, both are exactly the same.

[10] In our experiments, the constant *avgl* was fixed to the average length of documents in collection when a document is normal, but *avgl* was fixed to the average length of sample co-representation documents when a document is a co-representation of two documents. The sample set of co-representation documents was obtained from sets of top-retrieved documents in response to test queries used in Section 4.

4.2. Experimental setting

In this evaluation we used two different standard TREC collections, i.e., ROBUST and WT10G. Table 1 shows the basic statistics for each test collection, where *NumDocs* is the number of documents; *NumRels* is the total number of relevant documents for all topic set in each collection; *NumWords* is the total number of word occurrences in each collection; *TopicSet* is the range of topic numbers used for training and testing; and *Qrylen* is the average number of word occurrences in a query.

All experiments used the Lemur toolkit (version 4.12). We indexed English documents by performing standard preprocessing on queries and documents using Porter's stemmer, and we removed stopwords using the standard INQUERY stoplist (Allan, Connell, Croft, Feng, & Fisher, 2000). During all our evaluations, we only used words appearing in the query title.

There are two different types of baseline similarity. The first was term-based similarity, while the second was the existing TR similarity defined in Section 4.1.

The following is a summary of the baseline similarities that were compared in the experiments.

- **TSM**$_{VSM}$: Term-based similarity (TSM) based on a pivoted vector space model (i.e., Eq. (39)).
- **TR-QSSM**:Query-sensitive similarity (QSSM) proposed by Tombros and Rijsbergen (2001) (i.e., Eq. (32)).
- **TR-TSM + QSSM**: The combined metric of term-based similarity and query-sensitive similarity proposed by Tombros and Rijsbergen (2001) (i.e., Eq. (40)).

The following is a summary of our proposed methods that were compared in the experiments.

- **TSM**: Term-based similarity based on language modeling approaches (i.e., Eq. (27)).
- **CoR-QSSM**: The proposed (purely) query-sensitive similarity based on a co-relevance model without interpolation of term-based similarity (i.e., Eq. (26)).
- **CoR-TSM + QSSM**: The proposed combined similarity metric based on the co-relevance model (i.e., Eq. (9) with Eqs. (27) and (26)).

For notation simplicity, we often omit the prefix ''CoR'' when referring to our proposed method.

4.3. Evaluation measurement

Voorhees' NNT was used as an evaluation measure. In the original definition, the NNT-based evaluation measure is defined as a fraction given by the number of co-relevant documents from the top five similar documents to a given relevant document (Voorhees, 1985), which resembles the precision for five documents used in the ad hoc IR literature. NNT was also utilized in Tombros and Rijsbergen's experiment. NNT is basically an evaluation method for a ranked list, so we used mean average precision (MAP), P5, and P10 (precision for 5 and 10 documents, respectively) for NNT, which we refer to as NNT-MAP, NNT-P5, and NNT-P10, respectively.

For all comparisons, we applied paired *t*-test at 0.95 confidence level to report statistical significance.

4.3.1. Two types of NNT metrics: nonnormalized and normalized by query

Note that each query has a different number of relevant documents. Thus, if we merely use all relevant documents as a single entry, the resulting NNT result values (e.g., NNT-MAP or NNT-P5) could be highly influenced by queries with a large number of relevant documents. To compensate for this effect of queries with a large number of relevant documents, we also devise a normalized NNT metric, which we defined as the mean of all NNT results values over a set of queries, but not over a set of relevant documents. To present the formal definition of the nonnormalized and normalized NNT metrics, we assume that *sim* is a given similarity function to be evaluated, *Test* is the set of test queries, $\mathcal{R}(q)$ is set of relevant documents for q, and $NNT(sim, d, q)$ is the NNT result value of the similarity function *sim* for a relevant document d of query q using either NNT-P5, NNT-P10, or NNT-MAP. The nonnormalized version of NNT is defined as follows.

$$UnNormal(sim) = \frac{\sum_{q \in Test} \sum_{d \in \mathcal{R}(q)} NNT(sim, d, q)}{\sum_{q \in Test} |\mathcal{R}(q)|} \tag{41}$$

The normalized version of NNT is defined as follows.

Table 1
Statistics for each test collection.

Statistic	Robust	WT10G
NumDocs	528,156	1,692,096
NumWords	572,180	6,346,858
NumRels	17,412	5910
TopicSet	Q301–450 Q601–700	Q451–550
QryLen	2.58	2.56

$$Normal(sim) = \frac{1}{|Test|}\sum_{q\in Test}\sum_{d\in\mathcal{R}(q)}\frac{NNT(sim,d,q)}{|\mathcal{R}(q)|} \tag{42}$$

Before summing them, NNT result values are first normalized by the number of relevant documents for their query. To compute $NNT(sim,d,q)$ using the NNT-MAP, our evaluation was based on the top 1000 retrieved documents.

4.4. Parameter training

There are several parameters for each retrieval method, i.e., the pivot parameter s, the smoothing parameter μ, the interpolation parameter α (or β), and the prior probability $P(r)$. With TR similarity, we fixed s to 0.2 when computing term-based similarity, based on the suggestion of Singhal (2001). When computing query-sensitive similarities, we used a separate value of $s = 0.05$ because we observed that this setting had a better performance than $s = 0.2$. The same settings of s were also applied to the combined TR similarity in Eq. (40). In our proposed co-relevance-based similarity, we set μ to 1000 in all experiments, based on the suggestion of Lv and Zhai (2009).

Other parameters such as α and $P(r)$ were trained as follows. Given a test topic set consisting of 50 queries, each parameter was trained using other topic sets from the same test collection. For example, when Q301–350 in ROBUST was provided as a test set, a parameter was trained using Q351–450 and Q601–700 from the same ROBUST collection. When estimating $P(r)$, we used all relevant judgments from the pooled documents available in the training query set. To formally describe how we estimated $P(r)$, let $Train$ be the set of training queries, while $\mathcal{R}_{pool}(q)$ and $\mathcal{NR}_{pool}(q)$ are the sets of relevant and nonrelevant documents in the pooled documents selected by the pooling method, respectively. $P(r)$ is then estimated as follows.

$$P(r) = \frac{\sum_{q\in Train}|\mathcal{R}_{pool}(q)|}{\sum_{q\in Train}|\mathcal{R}_{pool}(q)| + \sum_{q\in Train}|\mathcal{NR}_{pool}(q)|} \tag{43}$$

where $|\mathcal{R}_{pool}(q)| + |\mathcal{NR}_{pool}(q)|$ is the number of pooled documents for query q.

4.5. Truncation when estimating the query-sensitive similarity

In our experiments, we applied truncation to restrict the set of reference documents when computing the query-sensitive similarity and term-based similarity. With truncation, we only referred to the top-retrieved T documents in response to q when computing $sim_{QSSM}(d,d',q)$ and the most similar M documents to d when computing $sim_{TSM}(d,d')$, as proposed by Tombros and Rijsbergen (2001). As a result, our final combined metric $sim_{TSM+QSSM}(d,d',q)$ used a maximum of $T + M$ documents. T and M were fixed at 1000 in all our experiments.

However, the resulting similarity with truncation would simply be zero when we only assign zeros to $sim_{QSSM}(d,d',q)$ and $sim_{TSM}(d,d')$, because $sim_{QSSM}(d,d',q)$ and $sim_{TSM}(d,d')$ are both considered as probabilities, in our method. To handle the issue of zero probability, we ''approximately'' assigned non-zero probabilities to similarities even in cases where a document did not appear in the top-retrieved documents or the most similar documents. To clearly demonstrate our method, suppose that d is a given document, $\mathcal{F}(q)$ is the set of top-retrieved T documents for query q using query-sensitive similarity $sim_{QSSM}(d,d',q)$, and $\mathcal{N}(d)$ is the set of the most similar M documents to d using term-based similarity $sim_{TSM}(d,d')$. There are three possible cases for d':

(1) $d' \in \mathcal{F}(q) \cap \mathcal{N}(d)$
(2) $d' \notin \mathcal{F}(q)$ but $d' \in \mathcal{N}(d)$
(3) $d' \in \mathcal{F}(q)$ but $d' \notin \mathcal{N}(d)$

Case 1 is a normal case where no approximation is necessary, so standard original score functions for d' were used for $sim_{TSM}(d,d')$ and $sim_{QSSM}(d,d',q)$. In cases 2 and 3, we used the following approximating assumptions.

- In case 2, we assume that document d' had no common terms with q.
- In case 3, we assume that document d' had no common terms with d.

These assumptions were applied to the final computation of our proposed similarity and TR similarity. With our combined query-sensitive similarity, the above assumption leads to the following formulae for $O(r|d',\tilde{q})$ and $O(r|d',\tilde{d})$ to use in cases 2 and 3, respectively.

$$\begin{aligned} O(r|d',\tilde{q}) &= \frac{P(r)}{P(\bar{r})}\lambda_{d'}^{|\tilde{q}|} = \frac{P(r)}{P(\bar{r})}\lambda_{d'} \\ O(r|d',\tilde{d}) &= \frac{P(r)}{P(\bar{r})}\lambda_{d'}^{|\tilde{d}|} = \frac{P(r)}{P(\bar{r})}\lambda_{d'} \end{aligned} \tag{44}$$

where $c(w,d')$ is assumed to be 0 for the query word w in q for case 2, and for the document word w in d for case 3.

It was simpler to apply the same assumptions to TR similarity. If no common matching term exists between the document and the query, the relevance score of the document to the query is zero, according to Eqs. (37), (35), and (30). Based on the above assumption, the two similarities $sim_{QSSM}(d,d',q)$ and $sim_{TSM}(d,d')$ are both set to 0 in cases 2 and 3, respectively.

4.6. Experimental results

Tables 2 and 3 show experimental results comparing term-based similarity and query-sensitive similarity using nonnormalized and normalized NNT result values as evaluation measures, respectively. Rows labeled **TR** indicate the results in setting the TR metric, while rows labeled **CoR** indicate results with our proposed co-relevance-based similarity metric. We use **TSM** and **QSSM** to refer to term-based and query-sensitive similarity metrics, respectively and use **TSM + QSSM** to indicate the combined metric of both similarities. The marks †, §, ‡, and ¶ indicate statistical significance of improvements over TR similarities; † over TR-based TSM; § over TR-based QSSM; ‡ over for two noncombined TR-based metrics (i.e., TSM and QSSM); ¶ over all three TR metrics (i.e., TSM, QSSM, and TSM + QSSM).

Both tables show that query-sensitive similarities led to significant improvements compared to term-based similarities for all three evaluation measures (MAP, P5, and P10), and for the TR metric and our metric. This result reconfirms the argument of Tombros and Rijsbergen (2001). As was shown by Tombros and Rijsbergen (2001), we also found that query-sensitive similarities alone showed improvements over term-based similarity and that the improvement with the combination was greater than that with the noncombined method.

From the two query-sensitive metrics (the proposed similarity and TR similarity), the proposed similarity metric consistently showed greater improvement over TR similarity with all test collections, and in all cases, these differences were statistically significant. We tested whether this improvement was merely due to the advantage of using a better retrieval model or the better estimation with query-sensitive similarity, by further evaluating the TR similarity in combination using two variants. (1) Rather than the pivot vector space model, we used top-retrieved documents based on Dirichlet prior smoothing and applied the same query-sensitive similarity as with TR similarity. (2) In addition to the first variant, we combined the resulting TR-based query-sensitive similarity with the term-based similarity from Dirichlet prior smoothing which uses Eq. (23).

Tables 4 and 5 show the results of the two variants of TR similarity applied with Dirichlet prior smoothing, comparing to the proposed combined similarity metric in the last row. The two variants are denoted as follows:

– **TR-QSSM**$_R$: Query-sensitive similarity (QSSM) proposed by Tombros and Rijsbergen (2001) (i.e., Eq. (32)), where the top-retrieved documents are chosen based on the language modeling approaches.

Table 2
Comparison of TSM and QSSM for Tombros and Rijsbergen's method and the proposed method based on probabilistic co-relevance on the nonnormalized NNT metrics. The marks †, §, ‡, and ¶ indicate statistical significance of improvements over TR-based TSM, TR-based QSSM, two noncombined TR-based metrics (i.e., TSM and QSSM), and all three TR metrics (i.e., TSM, QSSM, and TSM + QSSM), respectively. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM$_{VSM}$	0.1215	0.4961	0.3932	0.0852	0.4267	0.3108
	QSSM	0.1558†	0.3257	0.3114	0.1958†	0.4509†	0.4335†
	TSM + QSSM	0.2253‡	0.6502‡	0.5596‡	0.2385‡	0.6412‡	0.5521‡
CoR	TSM	0.1286†	0.5034‡	0.3994‡	0.0961†	0.4679‡	0.3436†
	QSSM	0.1958‡	0.5041‡	0.4669‡	0.2116‡	0.4662‡	0.4338†
	TSM + QSSM	**0.2497**¶	**0.6783**¶	**0.5997**¶	**0.2594**¶	**0.6488**¶	**0.5572**¶

Table 3
Comparison of TSM and QSSM for Tombros and Rijsbergen's method and the proposed method based on probabilistic co-relevance on the normalized NNT metrics. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM$_{VSM}$	0.1730	0.4590	0.3461	0.1697	0.3899	0.2643
	QSSM	0.1947	0.3483	0.3058	0.1749	0.2910	0.2562
	TSM + QSSM	0.2894‡	0.6150‡	0.5020‡	0.2891‡	0.5389‡	0.4222‡
CoR	TSM	0.1822†	0.4654‡	0.3507†	0.1807†	0.4175‡	0.2830†
	QSSM	0.2426‡	0.4859§	0.4325‡	0.1984§	0.3346§	0.2933§
	TSM + QSSM	**0.3182**¶	**0.6415**¶	**0.5379**¶	**0.3043**¶	**0.5596**‡	**0.4408**¶

Table 4
Further comparison of Tombros and Rijsbergen's method and the proposed similarity metric in the combined method on the nonnormalized NNT metrics. The marks †, §, ‡, and ¶ indicate statistical significance of improvements over TSM + QSSM, TR-based TSM_R + QSSM, TSM_R + QSSM_R, and 'all' three TR combined metrics (i.e., TSM + QSSM, TSM_R + QSSM, and TSM_R + QSSM_R), respectively. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM + QSSM	0.2253	0.6502	0.5596	0.2385	0.6412	0.5521
	TSM_R + QSSM	0.2304[†]	0.6365	0.5496	0.2429[†]	0.6503[†‡]	0.5599[†]
	TSM_R + QSSM_R	0.2353[†§]	0.6382[§]	0.5521	0.2477[†§]	0.6463[†]	0.5602[†]
CoR	TSM + QSSM	**0.2497**[¶]	**0.6783**[¶]	**0.5997**[¶]	**0.2594**[¶]	**0.6488**[†]	**0.5572**[†]

Table 5
Further comparison of Tombros and Rijsbergen's method and the proposed similarity metric in the combined method on the normalized NNT metrics. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM + QSSM	0.2894	0.6150	0.5020	0.2891	0.5389	0.4222
	TSM_R + QSSM	0.2959[†]	0.6034	0.4964	0.2912	0.5450[‡]	0.4244[‡]
	TSM_R + QSSM_R	0.2987[†§]	0.6028	0.4967	0.2915	0.5370	0.4181
CoR	TSM + QSSM	**0.3182**[¶]	**0.6415**[¶]	**0.5379**[¶]	**0.3043**[¶]	**0.5596**[‡]	**0.4408**[†‡]

- **TR-TSM_R + QSSM_R**: The combined similarity metric of term-based similarity with query-sensitive one proposed by Tombros and Rijsbergen (2001) (i.e., Eq. (40)), where the top-retrieved documents are chosen based on the language modeling approaches.

The results of the first and the second variants are presented in the sub-rows named TSM_R + QSSM and TSM_R + QSSM_R in the row named TR, respectively. Here, TSM_R is the term-based similarity using Eq. (27) based on Dirichlet-prior smoothing which is the same as that for our proposed combined similarity in Eq. (9). The marks †, §, ‡, and ¶ indicate statistical significance of improvements over TR similarities; † over TR-based TSM + QSSM; § over TR-based TSM_R + QSSM; ‡ over TSM_R + -QSSM_R; ¶ over all three TR combined metrics (i.e., TSM + QSSM, TSM_{DP} + QSSM, and TSM_R + QSSM_R). There was a small improvement in comparison with the original TSM + QSSM of TR similarity, but this was still weaker than the performance of the co-relevance based metric. Thus, these results provide evidence that the further improvement with probabilistic co-relevance compared with TR similarity was not merely due to replacing the pivoted vector space model with a better retrieval model. These results also suggest that compared to TR similarity, the use of probability co-relevance was advantageous as a metric for defining query-sensitive similarity.

4.7. Comparison with language modeling approaches based on Tombros and Rijsbergen's co-representation

The two variants of the TR similarity reported in tables still rely on the vector space model. To make a fairer comparison with our approach, we connect TR similarity to the relevance probability and fully use the language models for implementing the TR similarity for Eq. (9). To this end, we further consider the probabilistic extension of the Tombros and Rijsbergen's (2001) similarity, formulated as follows:

$$sim_{QSSM}(d, d', q) \approx P(r|d \otimes d', q) \tag{45}$$

which is referred to as the *co-representation relevance probability*, meaning the probability that the co-representation of *d* and *d'*, $d \otimes d'$, is relevant to query *q*. Employing the language model estimate for the relevance probability, we obtain the rewritten form of Eq. (45) as follows:

$$sim_{QSSM}(d, d', q) \approx \frac{\exp\left(S_q^{LM}(d \otimes d')\right)}{\exp\left(S_q^{LM}(d \otimes d')\right) + K \times \exp(|q|)} \tag{46}$$

Based on Eq. (46), we evaluated the following similarity metrics based on the co-representation.

- **TR-QSSM-LM**: Query-sensitive similarity (QSSM) presented proposed by Tombros and Rijsbergen (2001) (i.e., Eq. (46)) applied to the language modeling approaches ($P(r)$ was fixed to 0.01).
- **TR-TSM + QSSM-LM**: The combined similarity metric of term-based similarity with query-sensitive one presented proposed by Tombros and Rijsbergen (2001) (i.e., using Eq. (9) with Eqs. (27) and (46)) ($P(r)$ was fixed to 0.01).

Table 6
Comparison of TSM, QSSM-LM, and TSM + QSSM-LM using co-representation based co-relevance, and our proposed decomposition style method on the nonnormalized NNT metrics. The number of top-retrieved documents ($|\mathcal{F}(q)|$) was fixed at 1000. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM	0.1286	0.5034	0.3994	0.0961	0.4679	0.3436
	QSSM-LM	0.1566†	0.3079	0.3012	0.1870†	0.3804	0.3875†
	TSM + QSSM-LM	0.2270†‡	0.6293†‡	0.5420†‡	0.2386†‡	0.6068†‡	0.5213†‡
CoR	TSM + QSSM	**0.2505**¶	**0.6834**¶	**0.6020**¶	**0.2585**¶	**0.6479**¶	**0.5596**¶

Table 7
Comparison of TSM, QSSM-LM, and TSM + QSSM-LM using co-representation based co-relevance, and our proposed decomposition style method on the normalized NNT metrics. The number of top-retrieved documents ($|\mathcal{F}(q)|$) was fixed at 1000. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
TR	TSM	0.1822	0.4654	0.3507	0.1807	0.4175	0.2830
	QSSM-LM	0.1888	0.3236	0.2975	0.1609	0.2469	0.2213
	TSM + QSSM-LM	0.2900†‡	0.5894†‡	0.4858†‡	0.2859†‡	0.5257†‡	0.4073†‡
CoR	TSM + QSSM	**0.3177**¶	**0.6422**¶	**0.5371**¶	**0.3046**¶	**0.5548**¶	**0.4398**¶

Tables 6 and 7 report the NNT results of similarity metrics based on co-representation relevance probability, compared with term-based similarity (Eq. (27)) and our proposed similarity (Eq. (9) with Eqs. (27) and (26)). The marks †, ‡, and ¶ indicate statistical significance of improvements over other similarities; † over TSM (using Eq. (27)); ‡ over TR-QSSM-LM; ¶ all three metrics (i.e., TSM, TR-QSSM-LM, and TR-TSM + QSSM-LM).

Both tables show that the combined similarity with the co-representation relevance probability (i.e., TSM + QSSM-TR) significantly improves term-based similarity, and the improvements are statistically significant. However, their NNT results do not beat our decomposed style of estimation. Thus, the results again show that the high NNT values in our proposed similarity are not just the outcomes of applying an improved retrieval model. In addition, the results support that the independence assumption used in our approximation shows comparative performances in this task, despite its simplification.

4.8. Effect of P(r)

The previous experiments were based on the use of a $P(r)$ trained with judgment data, but we now present results with the automatic estimation of the $P(r)$ value using fixed prior probabilities. Tables 8 and 9 show NNT results for our co-relevance based similarities under various settings of $P(r)$. For comparison, the row named "$P(r)$ = trained" contains results where $P(r)$ was based on the training method. The row named "$P(r)\sim 0$" refer to results where $P(r)$ is very close to 0 (i.e., taking very large value for K), so the odd-based co-relevance model of Eq. (29) was used as the combined metric. Generally, performance did not change significantly over our range of $P(r)$, which was smaller than 0.5. Compared to the training method, we found that when $P(r)$ was large, such as 0.5 or 0.1, the performance was highly limited and showed no improvement over the training method. This suggests that a large $P(r)$ value negatively affects IR performance; so, a small $P(r)$ value is

Table 8
Effect of $P(r)$ on co-relevance based query-sensitive similarity TSM + QSSM – on the nonnormalized NNT metrics. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
CoR	$P(r)$ = trained	0.2497	0.6783	0.5997	**0.2594**	0.6488	**0.5572**
	$P(r) = 0.8$	0.2398	0.6574	0.5805	0.2499	0.5971	0.5259
	$P(r) = 0.5$	0.2419	0.6651	0.5875	0.2503	0.6026	0.5208
	$P(r) = 0.1$	0.2491	0.6744	0.5956	0.2580	0.6374	0.5490
	$P(r) = 0.01$	**0.2502**	**0.6858**	**0.6057**	0.2571	0.6433	0.5501
	$P(r) = 0.001$	0.2497	0.6836	0.6030	0.2567	**0.6496**	0.5525
	$P(r) = 0.0001$	0.2481	0.6779	0.5984	0.2557	0.6452	0.5491
	$P(r) = 0.00001$	0.2481	0.6779	0.5982	0.2554	0.6439	0.5482
	$P(r) = 0.000001$	0.2481	0.6779	0.5997	0.2553	0.6438	0.5480
	$P(r) \sim 0$ (odd)	0.2481	0.6779	0.5982	0.2553	0.6438	0.5480

Table 9
Effect of $P(r)$ on co-relevance based query-sensitive similarity TSM + QSSM – on the normalized NNT metrics. Bold faced number indicates the best performance in each test collection and evaluation measure.

Coll		Robust			WT10G		
NNT		MAP	P5	P10	MAP	P5	P10
CoR	$P(r)$ = trained	0.3182	0.6415	0.5379	0.3043	0.5596	**0.4408**
	$P(r)$ = 0.8	0.3011	0.6126	0.5091	0.2830	0.5094	0.4102
	$P(r)$ = 0.5	0.3052	0.6200	0.5168	0.2955	0.5282	0.4201
	$P(r)$ = 0.1	0.3163	0.6356	0.5326	0.3048	0.5548	0.4391
	$P(r)$ = 0.01	**0.3225**	0.6565	**0.5502**	0.3041	0.5548	0.4358
	$P(r)$ = 0.001	0.3231	**0.6591**	0.5502	**0.3086**	**0.5661**	0.4385
	$P(r)$ = 0.0001	0.3202	0.6509	0.5431	0.3067	0.5612	0.4353
	$P(r)$ = 0.00001	0.3202	0.6511	0.5430	0.3072	0.5602	0.4341
	$P(r)$ = 0.000001	0.3202	0.6512	0.5430	0.3071	0.5602	0.4338
	$P(r)\sim 0$ (odd)	0.3202	0.6512	0.5430	0.3071	0.5602	0.4338

Table 10
$P(r)$ for each topic set used in co-relevance-based query-sensitive similarity.

	Robust	WT10G
NumRels	17,733	5980
NumNonRels	304,436	134,490
$P(r)$	0.05504	0.04257

preferred. Relatively small values of $P(r)$, such as 0.01 or 0.001, occasionally led to slight improvements over the training method "$P(r)$ = trained," but the magnitude of the change was small.

Table 10 shows the estimated values of $P(r)$ based on training methods for each collection, using all topics as a training set, i.e., 250 queries for ROBUST and 100 queries for WT10G. When $P(r)$ was very small for both collections, this gave probabilities of 0.04 $\sim$ 0.06. The estimated values of $P(r)$ explain why the performances of "$P(r)$ = trained" in Tables 8 and 9 fell approximately between the cases of $P(r)$ = 0.1 and $P(r)$ = 0.01.

4.9. Sensitivity of α

We tested the effect of α in the combination method. Based on the previous results of Section 4.8, we set $P(r)$ at 0.01 because it delivered relatively good performance. Fig. 1 shows performance curves for NNT-MAP, NNT-P5, and NNT-P10 using normalized metrics, with the ROBUST and WT10G test collections. The curves were similar for both test collections, but absolute performance was different. With both collections, the best results were obtained when using α = 0.2. Of the three measures, the MAP curve was more stable than other metrics. Fig. 1 shows that with small alpha values (i.e., 0.2), the combination method always led to greater improvements when compared with both term-based and query-sensitive similarities.

4.10. Application: pseudo-relevance feedback – viewed as a method using the query-sensitive similarity

When setting our proposed co-relevance-based similarity, we observe an interesting equivalence between pseudo-relevance feedback and query-sensitive similarity-based re-ranking. In our view, pseudo-relevance feedback can be intuitively considered as a method for re-ranking document d based on the probability that d is co-relevant with the feedback documents.

To understand this equivalence, we suppose that the top-most retrieved document d_{fb} is selected as the source for the pseudo-relevance feedback. Given d_{fb} and q, all documents are re-ranked according to the $score_{FB}(d)$, which is a new score for the document d, which is defined as the co-relevance probability of d_{fb} and d in response to query q.

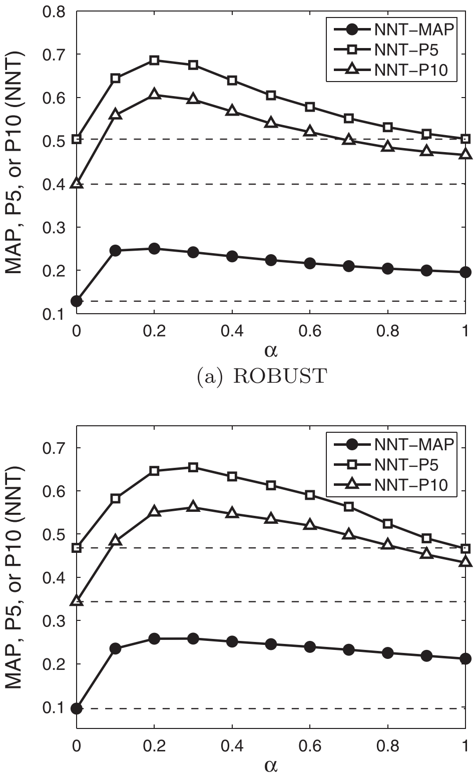
$$score_{FB}(d_{fb}, d, q) = P(corel|d_{fb}, d, q) \tag{47}$$

Using our proposed estimation of Eq. (9) for $P(corel|d_{fb}, d, q)$ and taking the logarithm, Eq. (47) is proportional to

$$score_{FB}(d_{fb}, d, q) \propto (1-\alpha)\ln P(r|d, Q = d_{fb}) + \alpha \ln P(r|d, q) + const. \tag{48}$$

where $const$ indicates a constant term that is independent of document d. To ensure that the relevance scores $score_{FB}(d)$ are highly discriminative among documents, we again assume that $P(r)$ is very small. Thus, taking K as a very large value, as in Section 3.3.4, this leads us to obtain the following formula using Eq. (29):

$$score_{FB}(d_{fb}, d, q) \propto (1-\alpha)\frac{S^{LM}_{d_{fb}}(d)}{|d_{fb}|} + \alpha \frac{S^{LM}_{q}(d)}{|q|} + const \tag{49}$$



(b) WT10G

Fig. 1. Performance curves (i.e., the nonnormalized NNT metrics) of the proposed query sensitive similarity method (TSM + QSSM), varying α.

In general, with more than one document, we could also select m documents $\mathcal{F} = \{d_{fb_1}, \ldots, d_{fb_m}\}$. A new score for the document d is defined as the co-relevance probability of the set $\mathcal{F}$ and d in response to query q:

$$score_{FB}(\mathcal{F}, d, q) = P(corel|\mathcal{F}, d, q) \tag{50}$$

Here, we introduce the notation $P(corel|\mathcal{F}, d, q)$ as a straightforward extension of $P(corel|d_{fb},d,q)$ to multiple documents, which represents the probability that all documents in $\mathcal{F}$ and d are co-relevant to q. Using the chain rule, Eq. (50) is decomposed to:

$$score_{FB}(\mathcal{F}, d, q) = P(corel|\mathcal{F}, q)P(r|d, \mathcal{F} = rel, q) \tag{51}$$

where $P(r|d, \mathcal{F} = rel, q)$ is the probability that d is relevant to the query q, given the condition that all the feedback documents in $\mathcal{F}$ are relevant. Note that $P(corel|\mathcal{F}, q)$ is constant for document d, so we do not focus on $P(corel|\mathcal{F}, q)$.

To estimate $P(r|d, \mathcal{F} = rel, q)$, we use the approximation introduced in Section 3.3 by dividing it into two different cases, i.e., (i) the relevance of d that is dependent on $\mathcal{F}$, and (ii) the relevant that is independent of $\mathcal{F}$.

In the first case, we use an additional assumption, i.e., the relevance of d that is conditionally independent of $\mathcal{F} - d_{fb_i}$, given the relevance of d_{fb_i}. Based on this assumption for each document d_{fb_i} in $\mathcal{F}$, we give the estimation for $P(r|d, \mathcal{F} = rel, q)$ as follows:

$$P_{DM_i}(r|d, \mathcal{F} = rel, q) \approx P(d \text{isrel}|d_{fb_i} \text{isrel}) \approx P(r|d, Q = d_{fb_i}) \tag{52}$$

In the second case, the independence assumption leads to the following simplifying estimation by removing the condition part q in $P(r|d, \mathcal{F} = rel, q)$ of Eq. (51):

$$P_{IM}(r|d, \mathcal{F} = rel, q) \approx P(r|d, q) \tag{53}$$

Thus, we have $m + 1$ estimations for $P(r|d, \mathcal{F} = rel, q)$ – $P(r|d, Q = d_{fb_1}) \cdots P(r|d, Q = d_{fb_m})$, and $P(r|d,q)$. Much like the derivation of the co-relevance probability used in Section 3.3, we take the geometric average of all estimations, which results in the following.

$$P(r|d, \mathcal{F} = rel, q) \approx \left(\prod_{d_{fb} \in \mathcal{F}} P(r|d, Q = d_{fb})^{\beta(d_{fb})} \right) P(r|d, q)^{1 - \sum_{d_{fb_i} \in \mathcal{F}} \beta(d_{fb_i})} \tag{54}$$

Using $\gamma(d_{fb}) = (1-\alpha)\beta(d_{fb})$, we further derive Eq. (55) to:

$$P(r|d,\mathcal{F}=rel,q) \approx \prod_{d_{fb}\in\mathcal{F}} \left(P(r|d,Q=d_{fb})^{1-\alpha} P(r|d,q)^{\alpha}\right)^{\gamma(d_{fb})} \tag{55}$$

By taking the logarithm of Eq. (55), $score_{FB}(\mathcal{F},d,q)(=\ln P(r|d,\mathcal{F}=rel,q))$ can now be expressed in terms of $score_{FB}(d_{fb_1},d,q),\cdots,score_{FB}(d_{fb_m},d,q)$ as follows:

$$score_{FB}(\mathcal{F},d,q) \propto \sum_{d_{fb}\in\mathcal{F}} \gamma(d_{fb}) \ln score_{FB}(d_{fb},d,q) = (1-\alpha)\left(\sum_{d_{fb}\in\mathcal{F}} \gamma(d_{fb}) \frac{S_{d_{fb}}(d)}{|d_{fb}|}\right) + \alpha\left(\frac{S_q(d)}{|q|}\right) + const \tag{56}$$

In a specific case, we consider the following formula to estimate the weight $\gamma(d_{fb})$.

$$\gamma(d_{fb}) = \frac{P(\tilde{q}|d_{fb},r)P(d_{fb}|r)}{\sum_{d'\in\mathcal{F}} P(\tilde{q}|d',r)P(d'|r)} = \frac{\exp\left(S^{LM}_{\tilde{q}}(d_{fb})\right)P(d_{fb}|r)}{\sum_{d'\in\mathcal{F}} \exp\left(S^{LM}_{\tilde{q}}(d')\right)P(d'|r)} \tag{57}$$

The resulting final score $score_{FB}(\mathcal{F},d,q)$ resembles the recomputed similarity scores using RM3 for pseudo-relevance feedback (Abdul-jaleel et al., 2004; Cartright, Allan, Lavrenko, & McGregor, 2010; Lv & Zhai, 2009; Lavrenko & Croft, 2001). The only difference between Eq. (56) and the similarity scores using RM3 is that RM3 uses a smoothed document language model $(1-\lambda_{d_{fb_i}})P_{ml}(w|d_{fb}) + \lambda_{d_{fb}}P(w|\mathcal{C})$ for d_{fb}, whereas our model uses an unsmoothed MLE document model $P_{ml}(w|d_{fb})$ for d_{fb}.

Thus, *pseudo-relevance feedback is a process for ranking documents in order, using the geometric average of co-relevance probabilities computed from feedback documents, i.e., using the weighted sum of query-sensitive similarities computed from feedback documents*. Thus, when viewed form the opposite perspective this suggests that query-sensitive similarity can be understood as a re-ranking method using pseudo-relevance feedback.

Provided query-sensitive similarity better satisfies the cluster hypothesis, this equivalence implies that the known effectiveness of pseudo-relevance feedback can be considered as further evidence supporting the cluster hypothesis. Croft (1980) also noted that pseudo-relevance feedback is another implementation that supports the cluster hypothesis. Indeed, one of the popular pseudo-relevance feedback methods, RM3, is a reformulation of interdocument relationships (Cartright et al., 2010). Smucker and Allan (2009) also defined query-sensitive similarity as a form of pseudo-relevance feedback, although they did not begin from the co-relevance probability.

5. Conclusion

This study revisited the query-sensitive similarity metric postulated by Tombros and Rijsbergen (2001) with the aim of developing a better form of metric that might satisfy the cluster hypothesis. Based on previous work, this study addressed the development of a probabilistic similarity metric and proposed the use of probabilistic co-relevance in a query-sensitive similarity metric. We first postulated a co-relevance-based similarity principle stating that the similarity between two documents should be proportional to the probability that they are co-relevant to a given query. We then decomposed co-relevance-based similarity into two components of relevance probabilities and described the proposed similarity metric in terms of a widely-used standard scoring function. Experimental results showed that the proposed query-sensitive similarity significantly improved upon existing term-based similarity in terms of evaluation metrics set in the context of Voorhees' NNT measure.

As a future work, it would be valuable to explore a discriminative model to estimate a co-relevance model based on a set of features obtained between a document and a query. This would develop a learning framework for the retrieval model, rather than our simplified decomposition. Data could be collected from a group of users, rather than a single user, allowing many relevant documents to be obtained, making a learnable mechanism more applicable in practice. These forms of collective feedback are available in realistic web search environments, so it would be valuable to investigate how to best use such implicit or explicit feedback and improve query-sensitive similarity.

References

Abdul-jaleel, N., Allan, J., Croft, W. B., Diaz, O., Larkey, L. & Li, X. (2004). Umass at trec 2004: Novelty and hard. In *Proceedings of the 13th text retrieval conference, TREC '04*.

Allan, J., Connell, M. E., Croft, W. B., Feng, F., Fisher, D. & Li X. (2000). Inquery and trec-9. In *Proceedings of the 9th text retrieval conference, TREC '00*.

Blanco, R. & Barreiro, A. (2008). Probabilistic document length priors for language models. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval, ECIR'08* (pp. 394–405).

Buckley, C., Salton, G., Allan, J. & Singhal, A. (1994). Automatic query expansion using smart. In *Proceedings of the 3rd text retrieval conference, TREC-3*.

Cartright, M.A., Allan, J., Lavrenko, V. & McGregor, A. (2010). Fast query expansion using approximations of relevance models. In *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10* (pp. 1573–1576).

Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems, 5*, 189–195.

El-Hamdouchi, A. & Willett, P. (1986). Hierarchic document classification using ward's clustering method. In *Proceedings of the 9th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '86* (pp. 149–156).

Fuhr, N., Lechtenfeld, M., Stein, B., & Gollub, T. (2011). The optimum clustering framework: Implementing the cluster hypothesis. *Information Retrieval, 15*(2), 93–115.

Goffman, W. (1968). An indirect method of information retrieval. *Information Storage and Retrieval, 4*, 361–373.

Griffiths, A., Robinson, L., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*(3), 175–205.

Hearst, M. A. & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '96* (pp. 76–84).

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the second European conference on research and advanced technology for digital libraries, ECDL '98* (pp. 569–584).

Hiemstra, D. (2001). *Using language models for information retrieval*. PhD thesis, University of Twente.

Jardine, N., & Rijsbergen, C. J. V. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval, 7*, 210–240.

Kurland, O. & Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 194–201).

Kurland, O. & Lee, L. (2005). Pagerank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05* (pp. 306–313).

Lafferty, J. & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01* (pp. 111–119).

Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. *Kluwer International Series on Information Retrieval – Language Modeling and Information Retrieval, 13*, 258.

Lavrenko, V. (2010). *A generative theory of relevance*. Incorporated: Springer Publishing Company.

Lavrenko, V. & Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01* (pp. 120–127).

Lavrenko, V., Allan, J., DeGuzman, E., Daniel, L., Pollard, V. & Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proceedings of the second international conference on human language technology research, HLT '02* (pp. 115–121).

Liu, X. & Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 186–193).

Losada, D. E. & Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval, 11*.

Lv, Y. & Zhai, C. (2009). A comparative study of methods for estimating query language models with pseudo feedback. In *Proceeding of the 18th ACM conference on information and knowledge management, CIKM '09* (pp. 1895–1898).

Miller, D. R. H., Leek, T. & Schwartz, R. M. (1999). A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '99* (pp 214–221).

Na, S. H., Kang, I. S., Roh, J. E., & Lee, J. H. (2007). An empirical study of query expansion and cluster-based retrieval in language modeling approach. *Information Processing and Management, 43*(2), 302–314.

Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '98* (pp. 275–281).

Rijsbergen, C. J. V. (1979). *Information retrieval*. Butterworths.

Rijsbergen, C. J. V., & Croft, W. B. (1975). Document clustering: An evaluation of some experiments with the cranfield 1400 collection. *Information Processing and Management*, 171–182.

Robertson, S. E. (1977). The probabilistic ranking principle in IR. *Journal of Documentation, 33*, 294–304.

Robertson, S. E. & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '94* (pp. 232–241).

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.

Roelleke, T. & Wang, J. (2006). A parallel derivation of probabilistic information retrieval models. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 107–114).

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of ACM, 18*(11), 613–620.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin, 24*(4), 35–43.

Singhal, A., Buckley, C. & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '96* (pp. 21–29).

Smucker, M.D. & Allan, J. (2005). *An investigation of dirichlet prior smoothings performance advantage*. Tech. rep., CIIR Technical Report IR-548. University of Massachusetts, Amherst.

Smucker, M. D. & Allan, J. (2009). A new measure of the cluster hypothesis. In *Proceedings of the 2nd international conference on theory of information retrieval: advances in information retrieval theory, ICTIR '09* (pp. 281–288).

Tao, T., Wang, X., Mei, Q. & Zhai, C. (2006). Language model information retrieval with document expansion. In *HLT-NAACL '06* (pp. 407–414).

Tombros, A. & Rijsbergen, C. J. V. (2001). Query-sensitive similarity measures for the calculation of interdocument relationships. In *Proceedings of the tenth international conference on Information and knowledge management, CIKM '01* (pp. 17–24).

Tombros, A., & Rijsbergen, C. J. V. (2004). Query-sensitive similarity measures for information retrieval. *Knowledge and Information Systems, 6*, 617–642.

Tombros, A., Villa, R., & Rijsbergen, C. J. V. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management, 38*, 559–582.

Voorhees, E. M. (1985). The cluster hypothesis revisited. In *Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '85* (pp. 188–196).

Wei, F., Li, W., Lu, Q. & He, Y. (2008). Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08* (pp. 283–290).

Willett, P. (1985). Query specific automatic document classification. *International Forum on Information and Documentation, 10*(2), 28–32.

Zhai, C. & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01* (pp. 334–342).

Zhou, Z. H. & Dai, H. B. (2006). Query-sensitive similarity measure for content-based image retrieval. In *Proceedings of the sixth international conference on data mining, ICDM '06* (pp. 1211–1215).