Navigating in the Dark: Modeling Uncertainty in Ad Hoc Retrieval Using Multiple Relevance Models

Natali Soskin, Oren Kurland, and Carmel Domshlak

Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel
natalis@tx.technion.ac.il,{kurland,dcarmel}@ie.technion.ac.il

Abstract. We develop a novel probabilistic approach to ad hoc retrieval that explicitly addresses the uncertainty about the information need underlying a given query. In doing so, we account for the special role of the corpus in the retrieval process. The derived retrieval method integrates multiple *relevance models* by using estimates of their *faithfulness* to the presumed information need. Empirical evaluation demonstrates the performance merits of the proposed approach.

Keywords: relevance models, ad hoc retrieval, faithfulness measures.

1 Introduction

The ad hoc retrieval task is to find documents relevant to an information need expressed by a query. However, it is often a hard challenge to infer what the underlying information need is, especially in the case of ambiguous queries.

We present a novel probabilistic framework to ad hoc retrieval that *explicitly* addresses the uncertainty about the information need expressed by a query. In doing so we account for two major factors that affect uncertainty, namely (1) the fact that the same query can be used to represent different information needs, and (2) the "nature" of the corpus upon which the search is performed. A case in point for the latter, a query for the car Jaguar used over the Web should better include the term "car", yet this term has no discriminative power in a portal dedicated to cars. The retrieval model that we derive integrates *multiple* relevance models [1,2], e.g., statistical language models that are presumed to generate terms in relevant documents. These relevance models potentially correspond to information needs that may underlie the query.

Our framework can be instantiated in various ways to yield specific retrieval algorithms, varying, for example, in the set of relevance models considered and in the *faithfulness* we attribute to each of them with respect to an information need presumably represented by the query.

To exemplify the practical potential of our framework, we take a pseudo feedback approach, and construct multiple relevance models based on documents sampled from an initially retrieved list. We then propose several faithfulness measures. Empirical evaluation demonstrates the performance merits of our methods with respect to using a single relevance model.

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 79–91, 2009.

2 Retrieval Framework

Taking a probabilistic approach to the task of ad hoc retrieval, our basic goal is to estimate the probability $p(d|q)$ that a given document d is relevant to a given query q. Since the relevance of a document should in fact be determined with respect to the information need I_q represented by q rather than with respect to q itself, it is important to reason about that information need within the process of estimating $p(d|q)$. The latter task is obviously challenging, because q can potentially represent different information needs, and, because in the ad hoc setting we usually do not have any information about I_q other than q.

Hence, while we assume that q communicates an (arbitrary complex, yet) *single* information need of the user, we should still strive to model and reason about our uncertainty on what that information need actually is.

Having this agenda in mind, we first consider the *generative assumption for relevance* [1,2] that states:

Assumption 1 (generative assumption). *Given information need I, there exists a relevance model R that generates the content in queries representing I and in documents relevant to I.*

The probabilistic graphical model representing this assumption is depicted in Fig. 1a. Henceforth, bold-faced and regular letters correspond to random variables and values of these random variables, respectively. The focus of prior work with respect to Assumption 1 was on estimating (some form of) a relevance model R by treating q as an observed sample from it; the documents are then ranked using (an estimate of) $p(d|R)$ [1]. However, as we argue next, the *estimation* of R touches only a part of the overall picture of the retrieval process.

We observe two "operational" aspects by which Assumption 1, as well as its practical realization described above, can be enhanced. The first aspect concerns the uncertainty about the information need underlying q. Prior work has coupled the information need with the relevance model, and addressed the uncertainty as an implicit part of estimating R. Thus, while the graphical model induces

$$p(d|q) = \sum_{I,R} p(d|R)p(R|I,q)p(I|q) \ ,$$

in practice, a single relevance model $R^* = \arg\max_R p(R|q)$ was *selected* by the virtue of choosing a specific estimation procedure for R, and then $p(d|R^*)$ served for $p(d|q)$. Note that doing so can conceptually be viewed as *replacing the true evidence q on $\mathbf{q}$ with a* de facto *evidence R^* on $\mathbf{R}$*, obtained by assuming, for example, (i) independence of $\mathbf{R}$ from $\mathbf{I}$ given q, and (ii) $p(\mathbf{I}|q)$ is a uniform distribution over those I that can represent q. In other words, in practice, the treatment of the uncertainty about the underlying information need is (implicitly) embodied in the specific choice of estimation technique for R^* rather than being reasoned about in the overall probabilistic model that $\mathbf{R}$ is a part of.

The second aspect is that of the context. Even if we were to know "exactly" what the information need is, the nature of the searched corpus C should have a

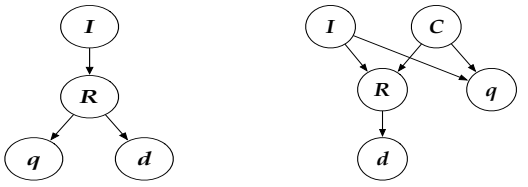

Fig. 1. Graphical-model representation of the (a) original, and (b) revised assumptions

significant impact on the way R is defined. For example, there might be language-specific issues (e.g., the terms "astronaut" and "cosmonaut" might need to be attributed with different levels of importance over English and Russian corpora, respectively). Moreover, a relevance model that can effectively discriminate (using some specific retrieval model) relevant documents from non-relevant ones over one corpus, might not be able to do so over a different corpus (e.g., recall the "Jaguar car" example from Sect. 1).

Given these two observations, we revise Assumption 1 as follows:

Assumption 2. *Given information need I and corpus C, there exists a relevance model R that generates documents relevant to I. Likewise, I and C determine the likelihood of a query q being selected to represent I in the context of C.*

Figure 1b depicts the corresponding graphical model. The major addition is modeling the dependence of $\mathbf{R}$ on the corpus $\mathbf{C}$. Note that the notion of corpus should be interpreted here not as a specific collection of documents, but rather as a *corpus characterization* (like "all documents on the Web" or "a collection of professional articles on various aspects of cardiology", etc.) Consequently, to represent the need for information about, for instance, "Jaguar cars", a relevance model R defined over the Web should better assign high importance to both the terms "Jaguar" and "car", while for R defined over a portal of cars, the term "car" should be assigned with low importance, if at all[1].

The second difference from Assumption 1 is that q is no longer assumed to be generated by R, but rather selected by a user to communicate her information need I in the context of C. For example, the user interested in a Jaguar car is less likely to use the single-term query "Jaguar" when searching over the Web, than when searching over a cars' portal. The reverse also holds, that is, the query "Jaguar" used over the Web should be considered by the search system to reflect a need for information with regard to both the car and the cat (potentially to varying degrees of importance) in lack of additional signal about the "true" need.

In the probabilistic model induced by Assumption 2, $\mathbf{R}$ still probabilistically depends on $\mathbf{q}$, but now via $\mathbf{I}$ and $\mathbf{C}$. That is, the observed query q reflects the latent information need I over C, while I and C determine the relevance model

[1] In practice, this could be done either implicitly (due to smoothing of document language models [3]), or explicitly [4].

R. We now turn to estimating $p(d|q)$ using Assumption 2. Specifically, we show that the indirect coupling between $\mathbf{R}$ and $\mathbf{q}$ via $\mathbf{I}$ and $\mathbf{C}$ allows for a direct reasoning about our uncertainty on the information need underlying q.

It is a fact that

$$p(d|q) = \sum_{C,I,R} p(d|R,I,C,q)p(I,C,R|q) \ ,$$

where the summation is over the universes of all corpora, information needs, and relevance models. Assumption 2 implies that the relevance of d to I can be determined based on R, and that R is uniquely determined given I and C. Thus,

$$p(d|q) = \frac{1}{p(q)} \sum_{C,I,R} p(d|R)p(R|I,C)p(I|q,C)p(q|C)p(C) \ ; \qquad (1)$$

this equation calls for a closer examination.

The first component of the summation term, $p(d|R)$, provides the common ground between Assumptions 1 and 2 — it is the probability that d is generated from R, i.e., that d is relevant to the information need represented by R. The other components involve the corpus C and are therefore specific to Assumption 2. First, $p(R|I,C)$ is either 0 or 1, depending on whether R is the one corresponding to the given I and C. Next, $p(I|q,C)$ is the probability that I is the information need communicated by q in the context of C. This component has an interesting property that the *entropy* of $p(\mathbf{I}|q,C)$ is essentially the *"query difficulty"* [5]. That is, the closer the distribution $p(\mathbf{I}|q,C)$ to uniform (i.e., higher entropy), the more difficult it is to infer the information need underlying q, and consequently, the harder it is to distinguish between relevant and non-relevant documents. Indeed, some estimates for query difficulty (a.k.a. *query performance*) rely on the connection between q and C [5]. Finally, the probability $p(q|C)$ could be viewed as referring to the potential of having information in C that corresponds on a surface-level to q. For example, if q is written in a different language than that used in C, then this correspondence will be low. As a result, while the appropriate relevance model R for C will be described in terms of the language used in C (e.g., a cross-lingual relevance model [6]), the probability of relevance, $p(d|q)$, will be lower than that for a query that uses the same language used in the corpus.

Deriving estimates for some of the probabilities in Eq. 1, specifically, $p(d|\mathbf{R})$ and $p(\mathbf{I}|q,\mathbf{C})$, is obviously a hard task. Therefore, we make the following pragmatic assumptions and estimation choices. First, we use the standard relevance *language model* approach [1] for the estimate $\hat{p}(d|R)$ — i.e., we use the probability that terms in d are generated by a statistical language model representing R. Second, since not query difficulty but uncertainty about the information need is of our focus in this work, for $p(\mathbf{I}|q,C)$ we adopt a very simple estimate $\hat{p}(\mathbf{I}|q,C)$ corresponding to a uniform distribution *only* over information needs that *can* be represented by q over C. Next, we use q as a proxy for those I it can represent as it is the only piece of information we have with respect to the underlying information need in lack of an informative prior over information needs and/or

additional user (relevance) feedback. Finally, we focus on the single-corpus search task (i.e., assume a single corpus C), and leave the multiple-corpora search task for future work, so as to arrive to the following rank equivalence:

$$\hat{p}(d|q) \stackrel{rank}{=} \sum_R \hat{p}(d|R)\hat{p}(R|q,C) \ . \quad (2)$$

It is important to note that while $p(R|I,C)$ is either 1 or 0, this is not the case for $\hat{p}(R|q,C)$ that results from using q as a proxy for I, as q can represent different information needs[2]. In what follows we treat $\hat{p}(R|q,C)$ as the probability that R corresponds to *some* information need I that is represented by q, where q is used to search information relevant to I over C.

2.1 Application

There are numerous ways of instantiating the ranking method presented in Eq. 2. However, to study the potential practical merits of the method, we need to make several implementation decisions. In what follows we present a possible set of such decisions, which constitutes only one example for how to use Eq. 2 to derive specific retrieval algorithms.

Our first task is to specify the set of relevance models to be utilized. There are various approaches for constructing relevance models (e.g, using documents [1], passages [7], document clusters [8], etc.). Here, we focus on utilizing documents for relevance-model construction.

Let $p(w|x)$ denote the probability assigned to term w by a (smoothed) unigram language model induced from text x. We use $\mathcal{D}_{init}^{[m]}$ ($\mathcal{D}_{init}$ in short) to denote the list of m documents d in the corpus C that yield the highest *query likelihood* $p(q|d) \stackrel{def}{=} \prod_{q_i} p(q_i|d)$; $\{q_i\}$ is the set of query terms. We define relevance model number 3 (RM3) [9] using the documents in $\mathcal{D}_{init}$:[3]

$$p(w|R) \stackrel{def}{=} \lambda p^{MLE}(w|q) + (1-\lambda) \sum_{d \in \mathcal{D}_{init}} p(w|d) \frac{\prod_{q_i} p(q_i|d)}{\sum_{d' \in \mathcal{D}_{init}} \prod_{q_i} p(q_i|d')} \ ; \quad (3)$$

$p^{MLE}(w|q)$ is the maximum likelihood estimate of w with respect to q; for $p(\cdot|d)$ we use a smoothed language model of d (further details in Sect. 4); λ is a free parameter.

[2] Since (i) R can represent different I's, and (ii) the estimate $\hat{p}(\mathbf{I}|q,C)$ is uniform over *only* those information needs that can be represented by q over C, $\hat{p}(\mathbf{R}|q,C)$ cannot be assumed to be uniformly distributed over the *entire* universe of relevance models.

[3] While RM3 assumes that q is generated from R this is not the case in Fig. 1b. We hasten to point out that using RM3 as an estimate for R here is only intended for performance-evaluation purposes, that is, to enable comparison of our paradigm that uses multiple relevance models with a state-of-the-art method that uses a single model. For full consistency with the graphical model, one could estimate R, for example, using a pseudo-feedback approach that only treats q as evidence for $\mathbf{I}$ (e.g., the state-of-the-art model-based feedback method [10]).

The list $\mathcal{D}_{init}$ often also contains non-relevant documents that may cause *query drift* [11] — i.e., shift between the information need underlying the query and that represented by the relevance model. Thus, as an alternative to using a single relevance model defined over $\mathcal{D}_{init}$, we define several relevance models that are constructed from documents sampled from $\mathcal{D}_{init}$. Specifically, we sample m sets of k documents (in Sect. 4 we compare random sampling [12] with cluster-based sampling [13]), and define over each set S a relevance model R_S using Eq. 3. Hopefully, some of the sampled sets will be composed of mainly relevant documents, or more generally, will faithfully reflect a "true" underlying information need. Naturally, the challenge, which we address below, is to quantify this faithfulness.

Our second order of business with respect to instantiating Eq. 2 is to estimate the probability of relevance model R generating the terms in document d, $\hat{p}(d|R)$. Some previous work [14] showed that in terms of retrieval effectiveness, using the cross-entropy between R and a language model induced from d is superior to estimating the probability that terms in d are generated from R. Thus, we use the complete-probability principle, and write:

$$\hat{p}(d|R) = \frac{\hat{p}(R|d)\hat{p}(d)}{\sum_{d'} \hat{p}(R|d')\hat{p}(d')} \ . \tag{4}$$

We assume a uniform prior distribution for documents, $\hat{p}(\mathbf{d})$, and use a cross-entropy-based measure: $\exp(-CE\left(p(\cdot|R) \mid\mid p(\cdot|d)\right))$=$\exp(\sum_w p(w|R) \log p(w|d))$ for the estimate $\hat{p}(R|d)$. We note that while this measure does not constitute a probability distribution, the resultant estimate $\hat{p}(\mathbf{d}|R)$ in Eq. 4 does.

2.2 "Faithfulness" of Relevance Models

The last and most important task towards instantiating Eq. 2 is devising the estimate $\hat{p}(\mathbf{R}|q, C)$ of the probability that a relevance model R represents an information need underlying q with respect to C.

The estimate for relevance model R_S is $\hat{p}(R_S|q, C) \stackrel{def}{=} \frac{F(R_S;q,C)}{\sum_{S'} F(R_{S'};q,C)}$, where $F(R_S; q, C)$ is a real-valued function quantifying the extent to which R_S presumably represents, or in other words, is faithful to, an information need underlying q with respect to C. The first faithfulness measure that we consider is the **uniform** distribution that represents the belief that all constructed relevance models in $\{R_S\}$ are faithful to the same extent:[4]

$$F_{uniform}(R_S; q, C) \stackrel{def}{=} 1 \ .$$

Next, the **constdoc** method estimates the faithfulness of R_S by the presumed percentage of relevant documents in S. Naturally, the more similar the constituent documents of S to the query are, the higher the estimate of this percentage should

[4] Note that there could be, and probably are, models in the universe of relevance models that are not in $\{R_S\}$ and that can faithfully represent the information need.

be. Following work on estimating the number of relevant documents in document clusters [15], we set:

$$F_{constdoc}(R_S; q, C) \stackrel{def}{=} \sqrt[|S|]{\prod_{d \in S} sim(q, d)} \ ,$$

where $sim(q, d) \stackrel{def}{=} \exp(-CE\left(p^{MLE}(\cdot|q) \mid\mid p(\cdot|d)\right)) = \sqrt[|q|]{p(q|d)}$ is d's normalized query likelihood [16]; CE is the cross-entropy and $|q|$ is q's length.[5]

Both faithfulness functions just described consider the corpus only indirectly. We therefore study the **clarity** method [5], which is based on the KL divergence between R_s and the corpus model:

$$\begin{aligned} F_{clarity}(R_S; q, C) &\stackrel{def}{=} \exp(KL\left(p(\cdot|R_S) \mid\mid p(\cdot|C)\right)) \\ &= \exp(\sum_w p(w|R_S) \log \frac{p(w|R_S)}{p^{MLE}(w|C)}) \ ; \end{aligned}$$

$p^{MLE}(w|C)$ is a maximum likelihood estimate of w with respect to C. The idea is that relevance models that are distant from the corpus model are "focused", and hence, are better candidates for representing a "coherent" information need [5]. Indeed, the value assigned by the clarity measure was shown to be somewhat correlated with the retrieval performance of the relevance model at hand [17].

The clarity measure does not consider (directly) the query for faithfulness estimation. The **drift** approach, in contrast, takes the query into account by measuring the divergence between the ranking induced by using R_S and that induced by using q. The idea is that the more distant the rankings are, the less faithful R_S is to the information need represented by q — i.e., the more chances there are for query drift [5,17]. The drift approach was shown to be effective for selecting a *single* relevance model from a set of candidates [17]. Formally, let L_q and L_{R_S} be the lists of 100 documents retrieved by using the original query q, and relevance model R_S, respectively. Let $p(w|L) \stackrel{def}{=} \beta \sum_{d_i \in L} p^{MLE}(w|d_i) + (1-\beta)p^{MLE}(w|C)$ be the language model induced from the document-list L; we set $\beta = 0.8$ [17]. The drift faithfulness measure is then:

$$\begin{aligned} F_{drift}(R_S; q, C) &\stackrel{def}{=} \exp(-CE\left(p(\cdot|L_q) \mid\mid p(\cdot|L_{R_S})\right)) \\ &= \exp(\sum_w p(w|L_q) \log p(w|L_{R_S})) \ . \end{aligned} \tag{5}$$

3 Related Work

Some previous work is conceptually similar to ours in that it addresses the uncertainty with respect to the information need by using multiple (manually-created) query representations [18,19,20]. However, no probabilistic framework

[5] Using the arithmetic mean of the document-query similarity values yields performance inferior to that of using the geometric mean.

was presented, and the "faithfulness" of a query representation to the underlying information need was not modeled.

Recent work [17] selects a single relevance model, using the clarity and drift measures, from a set of models constructed from the initial list $\mathcal{D}_{init}$. In Sect. 4 we demonstrate the merits of our approach with respect to this paradigm.

There are various methods — including document re-sampling as we use here [12,13] — for improving the retrieval effectiveness of relevance models (e.g.,[7,8,4,13]), and of query-expansion models that could be viewed as relevance models (e.g., [10,12]). These methods produce a single relevance model used for ranking, in contrast to our approach that uses multiple relevance models for ranking. However, our approach can potentially use these methods as it is not committed to a specific paradigm of relevance-model estimation.

4 Evaluation

4.1 Experimental Setup

We conducted experiments on four TREC data sets: (i) AP (disks 1-3, topics 51-150), (ii) SJMN (disk 3, topics 51-150), (iii) WSJ (disks 1-2, topics 151-200), and (iv) ROBUST (disks 4,5 (-CR), topics: 301-450, 601-700). Topics titles served as queries. We applied tokenization, Porter stemming, and stopword removal (using the INQUERY list) via the Lemur toolkit[6], which was also used for retrieval.

Unless otherwise specified, we use Dirichlet-smoothed unigram document language models with the smoothing parameter value set to 1000 [3]. The query-likelihood model [21], **QL**, in which document d is scored by $p(q|d) \stackrel{def}{=} \prod_{q_i} p(q_i|d)$ — i.e., the surface-level similarity between q and d — serves as a reference comparison to the algorithms we explore.

We use MAP (at cutoff 1000) and the precision of the top 5 documents (p@5) for performance evaluation. Statistically-significant differences of performance are determined using the two-tailed Wilcoxon test at the 95% confidence level.

An additional reference comparison to our methods is **RelModel** — a relevance model constructed from all documents in the initial list $\mathcal{D}_{init}$, which was retrieved using the query likelihood (QL) method; m, the number of documents in $\mathcal{D}_{init}$, is set to 50, as is the case for all other methods. The other free parameters that RelModel incorporates are set to values so as to optimize MAP. Specifically, λ, which controls the reliance on the original query model, is set to values in $\{0, 0.2, \ldots, 1\}$; the Jelinek-Mercer smoothing parameter of the language models of documents in $\mathcal{D}_{init}$ is chosen from $\{0.1, 0.2, \ldots, 1\}$; and, the number of terms used by the relevance model is set to values in $\{5, 10, 25, 50, 75, 100, 500\}$. The documents in the corpus are ranked in RelModel by the cross-entropy between the relevance model and their Dirichlet-smoothed language models.

To create the document sets $\{S\}$ upon which the multiple relevance models are constructed, we employed either nearest-neighbor-based clustering over $\mathcal{D}_{init}$ (with the KL-divergence as a similarity measure) in which each document served

[6] www.lemurproject.org

Table 1. Performance numbers. Best result in a column is boldfaced. Statistically significant differences with QL and RelModel are marked with 'l' and 'r', respectively

	AP		SJMN		WSJ		ROBUST	
	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5
QL	22.4	45.1	19.3	33.2	32.7	55.6	25.5	**48.2**
RelModel	28.9^{l}	50.7^{l}	24.1^{l}	38.4^{l}	38.7^{l}	**59.2**	27.6^{l}	46.9
uniform(rand)	28.6^{l}	50.7^{l}	23.4$^{l}_{r}$	38.4^{l}	38.3^{l}	58.8	27.1$^{l}_{r}$	47.7
uniform(clust)	29.3^{l}	52.7^{l}	24.5^{l}	39.8^{l}	39.2^{l}	57.2	26.9^{l}	46.0
constdoc(rand)	28.6^{l}	50.7^{l}	23.5^{l}	38.6^{l}	38.3^{l}	58.8	27.2$^{l}_{r}$	47.7
constdoc(clust)	**29.5**^{l}	52.7^{l}	24.5^{l}	39.0^{l}	39.4^{l}	56.8	**28.4**^{l}	47.9
clarity(rand)	28.7^{l}	50.5^{l}	23.6^{l}	39.0^{l}	38.3^{l}	58.0	27.0$^{l}_{r}$	47.7
clarity(clust)	29.3^{l}	52.9^{l}	24.6^{l}	39.6^{l}	**39.6**^{l}	58.0	27.1$^{l}_{r}$	44.7
drift(rand)	28.6^{l}	50.9^{l}	23.5^{l}	38.8^{l}	38.2^{l}	58.4	27.1$^{l}_{r}$	**48.2**
drift(clust)	29.3^{l}	**53.1**^{l}	**24.9**$^{l}_{r}$	**40.8**$^{l}_{r}$	39.2^{l}	58.0	27.7^{l}	47.9

as a basis for a cluster, or random selection from $\mathcal{D}_{init}$. In each case, 50 sets of k documents are used. Experiments with $k \in \{5, 10, 20\}$ under cluster-based selection showed that clusters of 5 and 10 documents yield relatively the same performance, while those of 20 documents yield inferior performance; hence, we set $k = 10$ in *all* tested models.

For computational reasons, we use each relevance model R to retrieve 1000 documents. Then, the lists retrieved by the multiple relevance models are *fused* using Eq. 2. Note that doing so simply amounts to setting $\hat{p}(R|d) = 0$ for all but 1000 documents d that yield the highest $\hat{p}(R|d)$. The other free parameters used to construct the multiple relevance models (λ, Jelinek-Mercer smoothing parameter, and number of terms) are set to the values chosen for RelModel, with which we compare our models as mentioned above. Hence, our multiple-relevance-models implementations are considerably *underoptimized* with respect to RelModel, as our goal is to focus on the underlying principles of our approach rather than engage in excessive tuning of parameters' values.

We use F(M) to denote a multiple-relevance-models implementation that uses the faithfulness measure F$\in$ {uniform, constdoc, clarity, drift} and the selection method M — either cluster-based (clust) or random-based (rand).

4.2 Results

Table 1 presents the performance numbers of our methods. Our first observation is that in most reference comparisons (corpus $\times$ evaluation measure) cluster-based selection of documents yields better performance than random-based selection. (Compare F(clust) with F(rand) rows.) This finding attests to the merit in constructing relevance models based on sets of similar documents that are potentially topically related. In addition, we note that both random-based and cluster-based implementations yield performance that is better (often to a statistically significant degree) than that of QL — the language model baseline.

The drift(clust) implementation yields, in general, the most effective performance among the implementations we consider. Thus, the divergence between the ranking induced by a relevance model and that induced by using the original

Table 2. Comparison with cluster-based document re-sampling (CBRSD) [13] for relevance-model construction. '>QL': percentage of queries for which the performance transcends that of QL. Boldface: best result in a column. Statistically significant differences with QL, RelModel, and CBRSD are marked with 'l', 'r', and 'c', respectively.

	AP				SJMN			
	MAP	>QL	p@5	>QL	MAP	>QL	p@5	>QL
QL	22.4	–	45.1	–	19.3	–	33.2	–
RelModel	28.9^{l}	**72.0**	50.7^{l}	34.0	24.1^{l}	67.0	38.4^{l}	31.0
CBRSD	$\mathbf{29.3}^{l}$	68.0	49.3	30.0	24.4^{l}	60.0	38.2^{l}	31.0
drift(clust)	$\mathbf{29.3}^{l}$	70.0	$\mathbf{53.1}^{l}_{c}$	**36.0**	$\mathbf{24.9}^{l}_{r}$	**68.0**	$\mathbf{40.8}^{l}_{r}$	**37.0**

	WSJ				ROBUST			
	MAP	>QL	p@5	>QL	MAP	>QL	p@5	>QL
QL	32.7	–	55.6	–	25.5	–	48.2	–
RelModel	38.7^{l}	**72.0**	**59.2**	**34.0**	27.6^{l}	61.2	46.9	25.2
CBRSD	$\mathbf{39.9}^{l}$	70.0	58.4	32.0	$\mathbf{30.7}^{l}$	**63.6**	**50.0**	**30.0**
drift(clust)	39.2^{l}	**72.0**	58.0	32.0	27.7^{l}_{c}	57.6	47.9	28.4

query, which is measured by the drift measure, seems to be a relatively effective estimate for the "faithfulness" of the relevance model to a presumed underlying information need. This finding is in accordance with a previous report about using drift to select a single relevance model from a set of models [17].

We can also see in Table 1 that all cluster-based implementations yield performance that is better in a majority of the relevant comparisons than that of RelModel, which constructs a single relevance model from all documents in $\mathcal{D}_{init}$. While the performance differences are, in general, not to a large scale, drift(clust), our best-performing method, outperforms RelModel to a statistically significant degree over SJMN for both MAP and p@5; also, there is only a single case (p@5 for WSJ) in which drift(clust) is outperformed by RelModel and the difference is not statistically significant. Recall that the performance of RelModel was optimized with respect to three free parameters, while that of our multiple-relevance-models was not optimized (except for the general choice of document-sets of size 10 for all implementations over all corpora). Thus, we view these results as gratifying, especially in light of the fact that parameters such as the number of terms are known to have considerable impact on the relevance-model performance. Furthermore, we note that RelModel can be used in Eq. 2 as one of the relevance models. Indeed, initial experiments with such implementation attest to the potential performance merits.

Performance Robustness. The relevance model, as other pseudo-feedback-based methods, suffers from a performance *robustness* problem [12,13]: for some queries the performance is worse than that of using only the original query (i.e., the QL method). Recent work [13] addresses this issue by constructing a relevance model using cluster-based document re-sampling (**CBRSD**) from $\mathcal{D}_{init}$ so as to "emphasize" documents with presumably high chances of relevance. We compare CBRSD — with re-sampling employed over the entire list $\mathcal{D}_{init}$ and the free

Table 3. Integrating multiple relevance models (our approach) vs. selecting a *single* (S-) relevance model [17] based on faithfulness measures. Boldface: best result in a column; 'l', 'r': statistically significant differences with QL and RelModel, respectively.

	AP		SJMN		WSJ		ROBUST	
	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5
QL	22.4	45.1	19.3	33.2	32.7	55.6	25.5	**48.2**
RelModel	28.9^{l}	50.7^{l}	24.1^{l}	38.4^{l}	38.7^{l}	59.2	27.6^{l}	46.9
S-constdoc(clust)	28.4^{l}	45.7	23.9^{l}	37.6^{l}	$\mathbf{40.0}^{l}$	$\mathbf{61.2}$	$\mathbf{28.8}^{l}$	45.9
constdoc(clust)	$\mathbf{29.5}^{l}$	52.7^{l}	24.5^{l}	39.0^{l}	39.4^{l}	56.8	28.4^{l}	47.9
S-clarity(clust)	27.4^{l}_{r}	47.3	23.6^{l}	35.8	37.8^{l}	57.6	24.2^{l}_{r}	35.5^{l}_{r}
clarity(clust)	29.3^{l}	52.9^{l}	24.6^{l}	39.6^{l}	39.6^{l}	58.0	27.1^{l}_{r}	44.7
S-drift(clust)	27.2^{l}_{r}	41.4_{r}	23.3^{l}	35.4_{r}	33.0_{r}	53.2	25.8_{r}	40.3^{l}_{r}
drift(clust)	29.3^{l}	$\mathbf{53.1}^{l}$	$\mathbf{24.9}^{l}_{r}$	$\mathbf{40.8}^{l}_{r}$	39.2^{l}	58.0	27.7^{l}	47.9

parameters set to the same values as those in our models and in RelModel — and drift(clust) in Table 2. We also report for both MAP and p@5 the percentage of queries (denoted ">QL") for which the performance transcends that of the QL method (i.e., performance robustness).

As we can see in Table 2 the performance of drift(clust) is in general better than that of CBRSD on AP and SJMN, while the reverse holds for WSJ and ROBUST. In addition, in most relevant comparisons the performance of drift(clust) is more robust than that of CBRSD. Moreover, while drift(clust) is more robust than RelModel in a majority of the comparisons, CBRSD is less robust than RelModel in most comparisons. Thus, we see that our approach of using multiple relevance models can help to improve performance robustness.

Comparison with Model Selection. As mentioned above, the clarity and drift faithfulness measures were used in previous work to select a *single* relevance model from a set of relevance models constructed by using document sampling from the initial list $\mathcal{D}_{init}$ [17]. Hence, in Table 3 we compare this model-selection paradigm (rows denoted with S-) with our approach that uses faithfulness measures to integrate models. (The uniform faithfulness measure does not constitute a selection criterion and is therefore not presented.) We can see that in most cases selecting a single relevance model yields performance that is inferior to that of our approach, and to that of RelModel. Thus, we conclude that there is merit in integrating multiple relevance models over selecting a single one.

5 Conclusion

We presented a novel probabilistic approach to ad hoc retrieval that *explicitly* addresses the uncertainty about the information need underlying a query. Our derived method integrates multiple relevance models by using their estimated *faithfulness* to the presumed information need. Empirical evaluation demonstrated the merits of our approach.

Acknowledgments. We thank the reviewers for helpful comments. This paper is based upon work supported in part by Google's and IBM's faculty research awards. Any opinions, findings and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the sponsors.

References

1. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of SIGIR, pp. 120–127 (2001)
2. Lavrenko, V.: A Generative Theory of Relevance, PhD thesis. University of Massachusetts Amherst (2004)
3. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of SIGIR, pp. 334–342 (2001)
4. Li, X., Croft, W.B.: Improving the robustness of relevance-based language models. Technical Report IR-401, Center for Intelligent Information Retrieval. University of Massachusetts (2005)
5. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Precision prediction based on ranked list coherence. Information Retrieval 9(6), 723–755 (2006)
6. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: Proceedings of SIGIR, pp. 175–182 (2002)
7. Liu, X., Croft, W.B.: Passage retrieval based on language models. In: Proceedings of CIKM, pp. 375–382 (2002)
8. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proceedings of SIGIR, pp. 186–193 (2004)
9. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMASS at TREC 2004 — novelty and hard. In: Proceedings of TREC-13, pp. 715–725 (2004)
10. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM, pp. 403–410 (2001)
11. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of SIGIR, pp. 206–214 (1998)
12. Collins-Thompson, K., Callan, J.: Estimation and use of uncertainty in pseudo-relevance feedback. In: Proceedings of SIGIR, pp. 303–310 (2007)
13. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: Proceedings of SIGIR, pp. 235–242 (2008)
14. Lavrenko, V., Croft, W.B.: Relevance models in information retrieval. In: [22], pp. 11–56.
15. Liu, X., Croft, W.B.: Evaluating text representations for retrieval of the best group of documents. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 454–462. Springer, Heidelberg (2008)
16. Lafferty, J.D., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: Proceedings of SIGIR, pp. 111–119 (2001)
17. Winaver, M., Kurland, O., Domshlak, C.: Towards robust query expansion: Model selection in the language model framework to retrieval. In: Proceedings of SIGIR, pp. 729–730 (2007)

18. Saracevic, T., Kantor, P.: A study of information seeking and retrieving. iii. searchers, searches, and overlap. Journal of the American Society for Information Science 39(3), 197–216 (1988)
19. Belkin, N.J., Cool, C., Croft, W.B., Callan, J.P.: The effect of multiple query representations on information retrieval system performance. In: Proceedings of SIGIR, pp. 339–346 (1993)
20. Lee, J.H.: Analyses of multiple evidence combination. In: Proceedings of SIGIR, pp. 267–276 (1997)
21. Song, F., Croft, W.B.: A general language model for information retrieval (poster abstract). In: Proceedings of SIGIR, pp. 279–280 (1999)
22. Croft, W.B., Lafferty, J. (eds.): Language Modeling for Information Retrieval. Information Retrieval Book Series, vol. 13. Kluwer, Dordrecht (2003)