Inf Retrieval J (2015) 18:359–378
DOI 10.1007/s10791-015-9257-z

Negative query generation: bridging the gap between query likelihood retrieval models and relevance

Yuanhua Lv[1] · **ChengXiang Zhai**[2]

Received: 16 December 2013 / Accepted: 28 May 2015 / Published online: 6 June 2015

Abstract The language modeling approach to information retrieval has recently attracted much attention. In the language modeling retrieval models, we can score and rank documents based on the query likelihood method. From the theoretical perspective, however, the justification of the existing (standard) query likelihood method based on the probability ranking principle requires an unrealistic assumption about the generation of a "negative query" from a document, which states that the probability that a user who dislikes a document would use a query does not depend on the particular document. This assumption enables ignoring the negative query generation so as to justify using the basic query likelihood method as a retrieval function. In reality, however, this assumption does not hold because a user who dislikes a document would more likely avoid using words in the document when posing a query. This suggests that the standard query likelihood function is a potentially non-optimal retrieval function. In this paper, we attempt to improve the standard language modeling retrieval models by bringing back the component of negative query generation. Specifically, we propose a general and efficient approach to estimate document-dependent probabilities of negative query generation based on the principle of maximum entropy, and derive a more complete query likelihood retrieval function that also contains the negative query generation component. In addition, we further develop a more general probabilistic distance retrieval method to naturally incorporate query language models, which covers the proposed query likelihood with negative query generation as its special case. The proposed approaches not only bridge the theoretic gap between the

A short version of this work has appeared as a short paper in Proceedings of CIKM'2012 (Lv and Zhai 2012).

✉ Yuanhua Lv
yuanhual@microsoft.com

ChengXiang Zhai
czhai@illinois.edu

[1] Microsoft Research, Redmond, WA 98052, USA

[2] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

standard language modeling retrieval models and the notion of relevance, but also improves the retrieval effectiveness with (almost) no additional computational cost.

Keywords Negative query generation · Query likelihood · Language model · Relevance · Probability ranking principle · Principle of maximum entropy

1 Introduction

The language modeling approach to information retrieval (Ponte and Croft 1998) has recently enjoyed much success for many different retrieval tasks (Ponte and Croft 1998; Xu and Croft 1999; Zhai and Lafferty 2001b; Lafferty and Zhai 2001; Xu et al. 2001; Lavrenko et al. 2002; Si et al. 2002; Zhang et al. 2002; Cronen-Townsend et al. 2002; Liu and Croft 2002; Zhai et al. 2003; Ogilvie and Callan 2003; Shen et al. 2005; Tan et al. 2006; Balog et al. 2006; Fang and Zhai 2007; Zhai 2008; Lv and Zhai 2009; Tsagkias et al. 2011). In the language modeling approach, we assume that a query is a sample drawn from a language model: given a query Q and a document D, we compute the likelihood of "generating" query Q with a document language model estimated based on document D. We can then rank documents based on the likelihood of generating the query, i.e., query likelihood.

Although the query likelihood retrieval function has performed well empirically, there was criticism about its theoretical foundation (Robertson and Hiemstra 2001; Sparck-Jones and Robertson 2001; Sparck-Jones et al. 2003). In particular, Sparck Jones questioned "where is relevance?" (Sparck-Jones and Robertson 2001). Responding to this criticism, Lafferty and Zhai (2002) showed that the basic query likelihood retrieval method can be justified based on the probability ranking principle (Robertson 1977) which is regarded as the theoretical foundation of retrieval models.

However, from the theoretical perspective, the justification of using the basic query likelihood as a retrieval function based on the probability ranking principle by Lafferty and Zhai (2002) requires an unrealistic assumption about the generation of a "negative query" from a document, which states that the probability that a user who dislikes a document would use a query does not depend on the particular document. This assumption enables ignoring the negative query generation in justifying using the standard query likelihood method as a retrieval function. With this assumption, the query likelihood can intuitively be explained as the probability that a user who likes a document would pose the query. In reality, however, this assumption does not hold because a user who dislikes a document would more likely avoid using words in the document when posing a query. This suggests that the basic query likelihood function is a potentially non-optimal retrieval function.

In order to bridge this theoretical gap between the basic query likelihood retrieval function and the notion of relevance, in this paper, we attempt to bring back the component of negative query generation.

A main challenge in estimating the negative query generation component is to develop a general method for any document with respect to any query. Our solution to this problem is to estimate the probability of negative query generation purely based on document D so as to make it possible to incorporate the negative query generation component when retrieving any document. Specifically, we exploit document D to infer the queries that a user would use to avoid retrieving D based on the intuition that such queries would not

likely have any information overlap with D. We then propose an effective approach to estimate probabilities of negative query generation based on the principle of maximum entropy (Jaynes 1957), which leads to a document-dependent negative query generation component that can be computed efficiently. Finally, we derive a more complete query likelihood retrieval function that also contains the negative query generation component, which essentially scores a document with respect to a query according to the ratio of the probability that a user who likes the document would pose the query to the probability that a user who dislikes the document would pose the query.

Similar to the standard query likelihood, a major deficiency of the proposed query likelihood with negative query generation is that it cannot easily incorporate query language models. To solve this problem, we further develop a more general probabilistic distance retrieval method, inspired by the development of the KL-divergence retrieval method (Lafferty and Zhai 2001). With this method, we first estimate a regular document language model, a regular query language model, and a "negative document language model" based on the probabilities of negative query generation, and we then score a document with respect to a query based on the relative KL-divergence between the query language model and the corresponding document language model and between the query language model and the corresponding negative document language model. With this probabilistic distance retrieval method, feedback can also be naturally cast as to improve the estimate of the query language model based on the feedback information. Interestingly, this probabilistic distance retrieval method covers the proposed query likelihood model with negative query generation as its special case.

Experiment results on several standard test collections show that the proposed query likelihood retrieval function with negative query generation improves the retrieval effectiveness significantly, especially in cases when queries are verbose, with (almost) no additional computational cost. And the proposed KL-divergence retrieval method with negative document language model also works more effectively than the standard KL-divergence retrieval method to handle query language models. Overall, the proposed approaches not only bridge the theoretic gap between the standard language modeling retrieval models and the probability ranking principle, but also improve retrieval effectiveness significantly.

Due to their effectiveness, efficiency, generality, and theoretical soundness, the proposed query likelihood method with negative query generation and KL-divergence retrieval method with negative document language model can potentially replace the standard query likelihood method and the standard KL-divergence method respectively in all retrieval applications.

2 Language modeling retrieval models

We first briefly review previous work related to the development and extensions of the language modeling retrieval models.

The query likelihood retrieval method was first introduced by Ponte and Croft (1998) and also independently explored by Miller et al. (1999) and Hiemstra (2001). In this method, given a query Q and a document D, we compute the likelihood of "generating" query Q with a model θ_D estimated based on document D, and then score and rank the document based on the likelihood of generating the query:

$$Score(D, Q) = p(Q|\theta_D) \tag{1}$$

The query generation can be based on any language model. Different models make different assumptions about term occurrences. So far, using a multinomial distribution (Miller et al. 1999; Hiemstra 2001; Zhai and Lafferty 2001b) for θ_D has been most popular and most successful, which is also adopted in our paper. However, several other choices have also been explored, including the multiple Bernoulli distribution (Ponte and Croft 1998; Metzler et al. 2004), the multiple Poisson distribution (Mei et al. 2007), and the Hypergeometric distribution (Tsagkias et al. 2011). With the multinomial distribution, the query likelihood is

$$p(Q|\theta_D) = \prod_w p(w|\theta_D)^{c(w,Q)} \tag{2}$$

where $c(w, Q)$ is the count of term w in query Q.

According to the maximum likelihood estimator, we have the following estimation of the document language model θ_D for the multinomial model:

$$p_{ml}(w|\theta_D) = \frac{c(w, D)}{|D|} \tag{3}$$

where $c(w, D)$ represents the count of term w in document D, and $|D|$ is the document length. The document language model θ_D needs to be smoothed to overcome the zero-probability problem, and an effective method is the Dirichlet prior smoothing (Zhai and Lafferty 2001b):

$$p(w|\theta_D) = \frac{|D|}{|D| + \mu} p_{ml}(w|D) + \frac{\mu}{|D| + \mu} p(w|C) \tag{4}$$

here $p(w|C)$ is the collection language model and is estimated as $p(w|C) = \frac{c(w,C)}{\sum_{w'} c(w',C)}$, where $c(w, C)$ indicates the count of term w in the whole collection C, and μ is a smoothing parameter (Dirichlet prior) which is usually set empirically. Smoothing plays two different roles in the query likelihood retrieval method (Zhai and Lafferty 2001b): one role is to assign non-zero probabilities to terms that are not observed in the document, and the other role is to weaken the effect of non-discriminative terms in the query to achieve an "IDF" effect.

Assuming the Dirichlet prior smoothing method, we can rewrite the query likelihood scoring function as follows (Hiemstra 2000; Zhai and Lafferty 2001b):

$$\begin{aligned} p(Q|\theta_D) &\overset{rank}{=} \log p(Q|\theta_D) \\ &= |Q| \log \frac{\mu}{|D| + \mu} + \sum_{w \in Q \cap D} c(w, Q) \log \left(1 + \frac{c(w, D)}{\mu p(w|C)}\right) \end{aligned} \tag{5}$$

where $|Q|$ represents query length. It shows that, although the query likelihood method is motivated in a different way than a traditional model such as the vector-space model, it tends to boil down to retrieval functions that implement retrieval heuristics (such as TF-IDF weighting and document length normalization) similar to those implemented in a traditional model (Hiemstra 2000; Zhai and Lafferty 2001b).

In the past decade, many more complex variants of the query likelihood method have been proposed for ad hoc retrieval. For example, n-gram (Song and Croft 1999), dependence language model (Gao et al. 2004), and positional language model (Lv and Zhai 2009) have been explored to go beyond the bag-of-word assumption; the query likelihood was also

extended as a translation model to allow inexact matching of semantically related words (Berger and Lafferty 1999); a full Bayesian query likelihood was studied to consider uncertainty of an estimation of θ_D (Zaragoza et al. 2003); parsimonious language models was proposed to improve the discrimination of language models (Hiemstra et al. 2004); cluster-based smoothing methods were evaluated for document-specific smoothing (Liu and Croft 2004; Wei and Croft 2006; Tao et al. 2006), etc. Although these extensions often outperform the basic query likelihood, they tend to incur significantly more computational cost.

A major deficiency of the query likelihood method is that it cannot easily incorporate query language models, making it hard to exploit relevance or pseudo-relevance feedback (PRF) in the language modeling approach (Zhai and Lafferty 2001a). To address this problem, a probabilistic distance model called Kullback–Leibler (KL) divergence retrieval method was proposed by Lafferty and Zhai (2001) to score a document based on the negative KL-divergence between the document language model and the query language model. The KL-divergence method can actually cover the query likelihood retrieval model as a special case when the query language model is estimated based on only the query. Moreover, the development of the KL-divergence retrieval model (Lafferty and Zhai 2001), which explicitly models both document and query language models, has attracted many efforts to propose effective PRF methods for improving the estimate of query language models, (e.g., Lafferty and Zhai 2001; Lavrenko and Croft 2001; Zhai and Lafferty 2001a; Kurland et al. 2005; Diaz and Metzler 2006; Collins-Thompson and Callan 2007; Lv and Zhai 2010).

The query likelihood method and the KL-divergence method have been shown to perform well for a variety of retrieval tasks, including ad-hoc retrieval (Ponte and Croft 1998; Zhai and Lafferty 2001b; Lafferty and Zhai 2001), cross-lingual information retrieval (Xu et al. 2001; Lavrenko et al. 2002), distributed information retrieval (Xu and Croft 1999; Si et al. 2002), structured document retrieval (Ogilvie and Callan 2003), personalized and context-sensitive search (Shen et al. 2005; Tan et al. 2006), modeling redundancy (Zhang et al. 2002), predicting query difficulty (Cronen-Townsend et al. 2002), expert finding (Balog et al. 2006; Fang and Zhai 2007), passage retrieval (Liu and Croft 2002; Lv and Zhai 2009), subtopic retrieval (Zhai et al. 2003), etc.

However, to the best of our knowledge, there has been no related work on the estimation of negative query generation, except one of our short conference papers (Lv and Zhai 2012). That paper has studied a query likelihood retrieval model with negative query generation. This paper is a more complete report of the work, and has also significantly extended the previous short paper with an additional contribution to enable query language models and relevance feedback with the negative query generation component.

3 Negative query generation

To better understand the retrieval foundation of the query likelihood method, Lafferty and Zhai (2002) provided a general relevance-based derivation of the query likelihood method. Formally, let random variables D and Q denote a document and a query, respectively. Let R be a binary random variable that indicates whether D is relevant to Q or not. Following Sparck-Jones et al. (2000), we will denote by ℓ ("like") and $\bar{\ell}$ ("not like") the value of the relevance variable. The probability ranking principle (Robertson 1977) provides a justification for ranking documents for a query based on the conditional probability of relevance, i.e., $p(R = \ell|D, Q)$. This is equivalent to ranking documents based on the odds ratio, which can be further transformed using Bayes' Rule:

$$O(R=\ell|Q,D)=\frac{p(R=\ell|Q,D)}{p(R=\bar{\ell}|Q,D)}\propto\frac{p(Q,D|R=\ell)}{p(Q,D|R=\bar{\ell})} \tag{6}$$

There are two different ways to decompose the joint probability $p(Q, D \mid R)$, corresponding to "document generation" and "query generation" respectively. With document generation $p(Q,D|R)=p(D|Q,R)p(Q|R)$, we have

$$O(R=\ell|Q,D)\propto\frac{p(D|Q,R=\ell)}{p(D|Q,R=\bar{\ell})} \tag{7}$$

Most probabilistic retrieval models (Robertson and Sparck-Jones 1976; Sparck-Jones et al. 2000; Fuhr 1992) are based on document generation.[1] Fuhr (1992) has provided in-depth discussions in this direction.

Query generation, $p(Q,D|R)=p(Q|D,R)p(D|R)$, is the focus of this paper. With query generation, we end up with the following ranking formula:

$$O(R=\ell|Q,D)\propto\frac{p(Q|D,R=\ell)p(R=\ell|D)}{p(Q|D,R=\bar{\ell})p(R=\bar{\ell}|D)} \tag{8}$$

in which, the term $p(R \mid D)$ can be interpreted as a prior of relevance on a document, which can be used to encode any bias on documents. Without such extra knowledge, we may assume that this term is the same across all the documents and obtain the following simplified ranking formula:

$$O(R=\ell|Q,D)\propto\frac{p(Q|D,R=\ell)}{p(Q|D,R=\bar{\ell})} \tag{9}$$

There are two components in this model:

- Positive query generation $p(Q|D,R=\ell)$ is the probability that a user who likes document D would pose query Q. One assumption is then made that this probability can be equivalent to the probability of generating query Q by drawing words from the document language model θ_D. With this assumption, the positive query generation essentially leads to the basic query likelihood $p(Q|\theta_D)$. And it has also been proved by Luk (2008) that the basic query likelihood retrieval function is "strict rank equivalent" to the positive query generation probability.
- "Negative" query generation $p(Q|D,R=\bar{\ell})$ is the probability that a user who dislikes a document D would use a query Q, or in other words, the probability that a user uses a query Q to avoid retrieving document D (thus the name, negative query generation). To give an example, suppose that a user is interested in papers related to IR models, but dislikes the paper he is reading right now (i.e., D) that is about non-probabilistic models. Then the user may formulate a query "probabilistic retrieval models" to attempt to exclude such papers in his/her search. Here "probabilistic retrieval models" is a negative query for D.

When we rank documents based on the query likelihood retrieval function proposed in (Ponte and Croft 1998), we essentially only use the first component, i.e., $p(Q|D,R=\ell)$

[1] It has been pointed out by Robertson (2005) that this document generation approach by Lafferty and Zhai (2003) is not theoretically equivalent to the classical probabilistic retrieval model (Robertson and Sparck-Jones 1976) due to their different event spaces. This issue, however, is out the scope of this work which focuses on the query generation approach, i.e., the language modeling retrieval model (Ponte and Croft 1998).

with the second component, i.e., negative query generation $p(Q|D, R = \bar{\ell})$ ignored. Thus, in order to justify using the basic query likelihood (i.e., the positive query generation component) alone as the ranking formula, an implicit assumption has to be made to exclude this negative query generation component, which states that the probability that a user who dislikes a document would use a query does not depend on the particular document (Lafferty and Zhai 2002), formally

$$p(Q|D, R = \bar{\ell}) = p(Q|R = \bar{\ell}) \tag{10}$$

This assumption enables ignoring the negative query generation in the derivation of the basic query likelihood retrieval function, leading to the following scoring formula:

$$O(R = \ell|Q, D) \propto p(Q|D, R = \ell) = p(Q|\theta_D) \tag{11}$$

Under this assumption, Lafferty and Zhai (2003) shown that ranking based on the query likelihood retrieval model is equivalent to ranking based on the probability of relevance. That is, the classical probabilistic retrieval model (Robertson and Sparck-Jones 1976) and the basic language modeling approach (Ponte and Croft 1998) are theoretically equivalent for ranking documents.[2]

In reality, however, the assumption of the negative query generation, as shown formally in Formula 10, does *not* hold because *a user who dislikes a document would more likely avoid using words in the document when posing a query*, suggesting that there is a theoretical gap between the standard query likelihood and the notion of relevance, and the standard query likelihood function is a potentially non-optimal retrieval function.

In the following section, we attempt to improve the basic query likelihood function by *estimating, rather than ignoring* the component of negative query generation $p(Q|D, R = \bar{\ell})$.

4 Language modeling retrieval models with negative query generation

4.1 Negative document language models

Given any document D in the collection, what would a user like if he/she does not like D? We assume that there exists a "complement" document $\bar{D}$, and that if a user does not like D, the user would like $\bar{D}$.[3] That is, when generating query Q, if a user does not like D, the user would randomly pick words from $\bar{D}$. Formally,

$$p(w|D, R = \bar{\ell}) = p(w|\theta_{\bar{D}}) \tag{12}$$

It is usually the case that such a document $\bar{D}$ does not really exist in the document collection. One can regard it as a virtual document that needs to be constructed. The challenge now lies in how to estimate a language model $\theta_{\bar{D}}$, which we refer to as the "negative document language model" of D. Note that the negative document language model in our paper is still a

[2] Although Robertson (2005) pointed out that Lafferty and Zhai (2003)'s conclusion may not be a valid general inference from the original probability ranking principle (Robertson 1977) due to their inconsistent event spaces, and Aly et al. (2014) further argued that the connection between the standard probability ranking principle and the language modeling approach may not be established on the level of probabilistic models, Lafferty and Zhai (2003)'s work, however, still presented a formal and widely-accepted way to connect the language modeling approach to the notion of "relevance" that could answer the question: "where is relevance?".

[3] Note that D is not a binary random variable. D and $\bar{D}$ are two separate variables both of which can take as value any single document, but the value of $\bar{D}$ depends on D.

document language model, which is completely different from the relevance model $p(w|R = \ell)$ (Lavrenko and Croft 2001) and the irrelevance model $p(w|R = \bar{\ell})$ (Wang et al. 2008) which are *query* language models that capture the probability of observing a word w relevant and non-relevant to a particular information need respectively.

Ideally we should use many actual queries by users who do not want to retrieve document D to estimate the probability $p(w|\theta_{\bar{D}})$. For example, if a user sees a document in search results but does not click on it, we may assume that he/she would dislike the document. Under this assumption, we can use all the queries from the users who "dislike" the document to approximate $\bar{D}$. However, in practice, only very few search results will be shown to users, and certainly there are always queries that we would not even have seen. That is, this estimation strategy will suffer from a serious data sparseness problem. Yet, as a general retrieval model, we argue that the proposed method must have some way to estimate $\theta_{\bar{D}}$ for any document with respect to any query.

To this end, one straightforward way is using the background language model $p(w|C)$ to approximate $p(w|\theta_{\bar{D}})$, based on the intuition that almost all other documents in the collection are complementary to D:

$$p(w|\theta_{\bar{D}}) = p(w|C) \tag{13}$$

With this estimate of $p(w|\theta_{\bar{D}})$, the negative query generation component does not affect the ranking of documents, because the probability of negative query generation will be constant for all documents: in some sense, this provides a justification for making the document independency assumption about the negative query generation component in the standard query likelihood method. However, the content of document D is ignored in this estimation. The question is whether we can leverage the content of document D to estimate a document-dependent negative query generation model $p(Q|\theta_{\bar{D}})$.

We are interested in estimating $p(w|\theta_{\bar{D}})$ in a general way based on the content of document D so as to make it possible to incorporate a document dependent negative query generation component when retrieving any document. Our idea is based on the intuition that if a user wants to avoid retrieving document D, he/she would more likely avoid using words in the document when posing a query. That is, the user would like a document $\bar{D}$ with little information overlap with D. Given only document D available, the sole constraint of $\bar{D}$ is that, if a word w occurs in D, i.e., $c(w, D) > 0$, this word should not occur in $\bar{D}$:

$$c(w, \bar{D}) = \begin{cases} 0 & \text{if } \mathrm{c(w, D)} > 0 \\ ? & \text{otherwise} \end{cases} \tag{14}$$

This leads to a $\bar{D}$ that contains a set of words that do not exist in D, but the frequency values of these words are unknown.

How to determine the frequency of a word in $\bar{D}$ if it does not occur in D? As the probability distribution of such a word is unknown, according to the *principle of maximum entropy* (Jaynes 1957), each word occurring in $\bar{D}$ should have the same frequency $\delta > 0$, which maximizes the information entropy under the only prior data D. That is, $\bar{D}$ contains a set of words that are complementary to D in the universe word space (i.e., the whole word vocabulary V) with the same frequency δ. Formally,

$$c(w, \bar{D}) = \begin{cases} 0 & \text{if } \mathrm{c(w, D)} > 0 \\ \delta & \text{otherwise} \end{cases} \tag{15}$$

According to the maximum likelihood estimator, we have the following estimation of the document language model $\theta_{\bar{D}}$ for the multinomial model:

$$p_{ml}(w|\theta_{\bar{D}}) = \frac{c(w,\bar{D})}{|\bar{D}|} \tag{16}$$

where $|\bar{D}|$ is the "document" length of $\bar{D}$, which can be computed by aggregating frequencies of all words occurring in $\bar{D}$. Because the number of unique words not in $\bar{D}$ (i.e., appearing in D) is usually much smaller than the number of unique words in the whole document collection C (i.e., $|V|$), the number of unique words in $\bar{D}$ is thus approximately the same as $|V|$. That is

$$|\bar{D}| = \sum_{w \in V} c(w,\bar{D}) \approx \delta |V| \tag{17}$$

Due to the existence of zero probabilities, $p_{ml}(w|\theta_{\bar{D}})$ needs smoothing. Following the estimation of regular document language models, we also choose the Dirichlet prior smoothing method due to its effectiveness in information retrieval (Zhai and Lafferty 2001b). Formally,

$$p(w|\theta_{\bar{D}}) = \frac{\delta|V|}{\delta|V|+\mu} p_{ml}(w|\theta_{\bar{D}}) + \frac{\mu}{\delta|V|+\mu} p(w|C) \tag{18}$$

where μ is the Dirichlet prior. Since the influence of μ can be absorbed into variable δ, we thus set it simply to the same Dirichlet prior value as used for smoothing the regular document language model (see Eq. 4).

4.2 Query likelihood with negative query generation

Now we can bring back the negative query generation component to the query generation process based on the probability ranking principle:

$$\begin{aligned} O(R=\ell|Q,D) &\overset{rank}{=} \log \frac{p(Q|D,R=\ell)}{p(Q|D,R=\bar{\ell})} \\ &= \log p(Q|D,R=\ell) - \log p(Q|D,R=\bar{\ell}) \\ &= \log p(Q|\theta_D) - \log p(Q|\theta_{\bar{D}}) \end{aligned} \tag{19}$$

where the negative query loglikelihood $\log p(Q|\theta_{\bar{D}})$ can be further written as

$$\begin{aligned} &\log p(Q|\theta_{\bar{D}}) \\ &= \sum_{w \in Q} c(w,Q) \log p(w|\theta_{\bar{D}}) \\ &= \sum_{w \in Q \cap D} c(w,Q) \log \left(\frac{\mu \cdot p(w|C)}{\delta|V|+\mu} \right) + \sum_{w \in Q, w \notin D} c(w,Q) \log \left(\frac{\delta}{\delta|V|+\mu} + \frac{\mu \cdot p(w|C)}{\delta|V|+\mu} \right) \\ &= \sum_{w \in Q \cap D} c(w,Q) \log \left(\frac{\mu \cdot p(w|C)}{\delta|V|+\mu} \right) - \sum_{w \in Q \cap D} c(w,Q) \log \left(\frac{\delta + \mu \cdot p(w|C)}{\delta|V|+\mu} \right) \\ &+ \underbrace{\sum_{w \in Q} c(w,Q) \log \left(\frac{\delta + \mu \cdot p(w|C)}{\delta|V|+\mu} \right)}_{\text{document independent constant}} \\ &\overset{rank}{=} - \sum_{w \in Q \cap D} c(w,Q) \log \left(1 + \frac{\delta}{\mu p(w|C)} \right) \end{aligned} \tag{20}$$

Plugging Eqs. 5 and 20 into Eq. 19, we finally obtain a more complete query likelihood retrieval function that also contains the negative query generation component:

$$O(R=\ell|Q,D) \stackrel{rank}{=} |Q|\log\frac{\mu}{|D|+\mu}+\sum_{w\in Q\cap D}c(w,Q)\left[\log\left(1+\frac{c(w,D)}{\mu p(w|C)}\right)+\log\left(1+\frac{\delta}{\mu p(w|C)}\right)\right] \quad (21)$$

Comparing Formula 21 with the standard query likelihood in Formula 5, we can see that our new retrieval function essentially introduces a novel component $\log\left(1+\frac{\delta}{\mu p(w|C)}\right)$ to reward the matching of a query term, and it rewards more the matching of a more discriminative query term, which not only intuitively makes sense, but also provides a natural way to incorporate IDF weighting to query likelihood, which has so far only been possible through a second-stage smoothing step (Hiemstra 2000; Zhai and Lafferty 2002). Note that when we set $\delta = 0$, the proposed retrieval function degenerates to the standard query likelihood function.

Moreover, this new component will not change the relative score of two documents if they match the same number of unique query terms, but it will change the relative score of two documents if one matches a query term while the other does not. In this sense, we would hypothesize that (1) the proposed new retrieval function may not affect the standard query likelihood ranking of top result documents too much, as the top result documents tend more likely to match all the query terms, and (2) the proposed retrieval function would influence the standard query likelihood ranking more significantly for verbose queries, because the result documents tend to easily miss some query terms when a query is verbose. Both hypotheses are confirmed in our experiments.

Furthermore, since this new component we introduced is a *term-dependent constant*, the proposed new retrieval function only incurs $\mathcal{O}(|Q|)$ additional computation cost as compared to the standard query likelihood function, where $|Q|$ is the number of query terms. This can be certainly ignored.

Interestingly, the developed retrieval function based on query likelihood with negative query generation (Formula 21) leads to the same ranking formula as derived by lower-bounding term frequency normalization in the standard query likelihood method (Lv and Zhai 2011). However, the derivation of the ranking formula there is based on a heuristic approach; in this work, we show that the heuristic derivation is actually consistent with the probabilistic framework of the query likelihood method. In other words, query likelihood with negative query generation essentially provides a probabilistic interpretation for the heuristic method of lower-bounding term frequency normalization in the standard query likelihood method; meanwhile, this connection can also justify why the component of negative query generation is necessary to improve the standard query likelihood retrieval function.

4.3 KL-divergence retrieval method with negative query generation

Estimating query language models using relevance or PRF is an important technique to improve retrieval accuracy. However, similar to the standard query likelihood, a major deficiency of the proposed query likelihood with negative query generation is that it cannot easily incorporate query language models (Zhai and Lafferty 2001a). To address this problem, we further develop a more general probabilistic distance retrieval method to explicitly incorporate the query language model, inspired by the development of the KL-divergence retrieval method (Lafferty and Zhai 2001).

The key idea is that, the closer the document language model is to the query language model and the farther away the corresponding negative document language model is from the query language model, the higher the document would be ranked. Specifically, we choose the KL-divergence to measure the distance between two language models; given the regular document language model θ_D, the regular query language model θ_Q, and the proposed negative document language model $\theta_{\bar{D}}$ (Eq. 18), we score a document D with respect to a query Q based on the difference of two KL-divergence values: one is the KL-divergence between the query language model θ_Q and the regular document language model θ_D, and the other KL-divergence between θ_Q and the negative document language model $\theta_{\bar{D}}$. Formally,

$$\begin{aligned} Score(D,Q) &= D(\theta_Q||\theta_{\bar{D}}) - D(\theta_Q||\theta_D) \\ &= \sum_{w\in V} p(w|\theta_Q)\log\frac{p(w|\theta_Q)}{p(w|\theta_{\bar{D}})} - \sum_{w\in V} p(w|\theta_Q)\log\frac{p(w|\theta_Q)}{p(w|\theta_D)} \\ &= \sum_{w\in V} p(w|\theta_Q)\log\frac{p(w|\theta_D)}{p(w|\theta_{\bar{D}})} \end{aligned} \tag{22}$$

Moreover, it is easy to show that this probabilistic distance retrieval method covers the query likelihood with negative query generation as its special case when we use the actual/raw query word distribution to estimate θ_Q, i.e.,

$$p(w|\theta_Q) = \frac{c(w,Q)}{|Q|} \tag{23}$$

Indeed, with such an estimate, we have:

$$\begin{aligned} Score(D,Q) &= \sum_{w\in V} \frac{c(w,Q)}{|Q|}\log\frac{p(w|\theta_D)}{p(w|\theta_{\bar{D}})} \\ &\overset{rank}{=} \sum_{w\in V} c(w,Q)\log\frac{p(w|\theta_D)}{p(w|\theta_{\bar{D}})} \\ &= \log p(Q|\theta_D) - \log p(Q|\theta_{\bar{D}}) \end{aligned} \tag{24}$$

In this regard, the proposed probabilistic distance retrieval method is not only an extended KL-divergence model with a negative document language model but also a generalization of the proposed query likelihood with a negative query generation component. Although the proposed probabilistic distance retrieval method (Formula 22) and the basic query likelihood with a negative query generation (Formula 21) rank documents equally for the actual/initial query, as shown above, we would like, however, to emphasize that the major advantage of the former over the latter is that the former is also able to deal with a query language model $p(w|\theta_Q)$ that may be estimated using advanced query modeling techniques, such as PRF (Zhai and Lafferty 2001a; Lv and Zhai 2010).

5 Experiments

5.1 Testing collections and evaluation

We use four TREC collections: WT2G, WT10G, Terabyte, and Robust04, which represent different sizes and genre of text collections. WT2G, WT10G, and Terabyte are small,

medium, and large Web collections respectively. Robust04 is a representative news dataset. We use the Lemur toolkit and the Indri search engine[4] to carry out our experiments. For all the datasets, the preprocessing of documents and queries is minimum, involving only Porter's stemming. An overview of the involved query topics, the average length of short/verbose queries, the total number of relevance judgments, the total number of documents, the average document length, and the standard deviation of document length in each collection are shown in Table 1. We test three types of queries:

- Short queries, which are taken from the title field of the TREC topics.
- Verbose queries, which are taken from the description field of the TREC topics.
- Automatically expanded queries, which are generated using PRF techniques.

We employ a twofold cross-validation for parameter tuning, where the query topics are split into even and odd numbered topics as the two folds. Specifically, the parameters of each method are trained on even (odd) numbered topics and tested on the odd (even) numbered topics. The performance is then measured by combining both test sets. The top-ranked 1000 documents for each run are compared in terms of their mean average precisions (MAP), which also serves as the objective function for parameter training. In addition, the precision at top-10 documents (P@10) and the recall are also considered. The major goals of the experiments are to answer the following questions:

1. If the proposed negative query generation component can work well for improving the standard query likelihood method and the KL-divergence method?
2. How do the proposed techniques perform on different types of queries, such as short (keyword) queries, verbose queries, and automatically expanded queries?

5.2 Performance comparison

We first compare the effectiveness of the standard language modeling approach (labeled as LM) (Ponte and Croft 1998; Zhai and Lafferty 2001b) and the proposed language modeling approach with negative query generation (labeled as XLM) for both short and verbose queries without using feedback. Note that without using feedback, the query likelihood method with negative query generation (Sect. 4.2) and the KL-divergence method with negative query generation (Sect. 4.3) are rank equivalent. LM has one free parameter μ (for smoothing the standard document language model in Eq. 4) and XLM has two free parameters μ (for smoothing both the standard document language model in Eq. 4 and the negative document language model in Eq. 18) and δ. We use cross validation to train both μ and δ for XLM, and μ for LM.

We report the comparison results in Table 2. The results demonstrate that XLM outperforms LM consistently in terms of MAP and recall (#Rel) and also achieves better P@10 scores than LM in most cases. The MAP improvements of XLM over LM are significant in most cases. These results show that bringing back the negative query generation component is able to improve retrieval performance, and that the proposed approach to the estimation of negative query generation probabilities works effectively. Although XLM achieves better or comparable p@10 scores as compared to LM, their score differences are often minor; this verifies our first hypothesis in Sect. 4 that the negative query generation component may not influence too much the top-ranked result documents.

[4] http://www.lemurproject.org/.

Table 1 Document set characteristic

	Terabyte	WT10G	Robust04	WT2G
Queries	701–850	451–550	301–450, 601–700	401–450
#qry(with qrel)	149	100	249	50
avg(ql_short)	3.13	4.24	2.74	2.46
avg(ql_verb)	11.55	11.61	15.47	13.86
#total_qrel	28,640	5981	17,412	2279
#documents	25,205k	1692k	528k	247k
avdl	949	611	481	1056
std(dl)/avdl	2.63	2.31	1.19	2.14

Table 2 Comparison of the standard language modeling approach (LM) and the proposed language modeling approach with negative query generation (XLM) on short queries, verbose queries, and automatically expanded queries using PRF

We do significance test on MAP, where superscripts 1/2/3/4 indicate that the corresponding improvement is significant at the 0.05/0.02/0.01/0.001 level using the Wilcoxon non-directional test. "#Rel" stands for the number of total relevant documents retrieved

Bold font highlights the MAP improvements of XLM over LM

Dataset	Query type	Method	MAP	P@10	#Rel
WT2G	Short	LM	0.3088	0.4600	1905
		XLM	**0.3187**^{3}	0.4620	1920
	Verbose	LM	0.2742	0.4000	1837
		XLM	**0.2871**^{2}	0.4100	1854
	PRF	LM	0.3385	0.4880	1915
		XLM	**0.3474**^{2}	0.4800	1964
WT10G	Short	LM	0.1930	0.2796	3812
		XLM	**0.1961**	0.2807	3852
	Verbose	LM	0.1790	0.3150	3816
		XLM	**0.1871**^{3}	0.3140	3975
	PRF	LM	0.2205	0.3071	3809
		XLM	**0.2245**^{1}	0.3031	3931
Terabyte	Short	LM	0.2921	0.5463	19,391
		XLM	**0.2936**^{3}	0.5503	19,404
	Verbose	LM	0.2112	0.4718	14,468
		XLM	**0.2143**^{1}	0.4718	14,734
	PRF	LM	0.3278	0.5791	19,896
		XLM	**0.3339**^{3}	0.5858	20,075
Robust04	Short	LM	0.2521	0.4225	10,260
		XLM	**0.2530**^{1}	0.4229	10,244
	Verbose	LM	0.2329	0.3968	9344
		XLM	**0.2440**^{4}	0.3992	9372
	PRF	LM	0.2788	0.4382	10,741
		XLM	**0.2797**	0.4422	10,814

So far we have shown that the proposed language modeling approach with negative query generation outperforms the standard language modeling approach when basic queries are used without exploiting any feedback information. Next, we turn to examine if negative query generation can also benefit the automatically expanded queries using PRF. Specifically, we use the standard language modeling method to do an initial retrieval for each short query, and then apply the positional-relevance model (PRM1) (Lv and Zhai

2010), which is a state-of-the-art PRF algorithm, to estimate an improved query language model. After that we use the obtained query language model to rank documents using the proposed KL-divergence retrieval method with a negative document language model (Eq. 22); in addition, we also use the *same* query language model in the standard KL-divergence retrieval method (Lafferty and Zhai 2001) as our baseline.[5] The parameters μ and δ in the scoring functions are trained using cross validation.

We present the results in Table 2. It shows that XLM consistently outperforms LM, suggesting that the negative query generation component also works well for PRF. This demonstrates that, through bringing back the negative query generation, we have developed a general retrieval method that works well for different types of queries.

Comparing short with verbose queries, we observe that XLM generally improves more on verbose queries than on short queries. In particular, the MAP improvements on WT2G, WT10G, and Robust04 collections are as high as 5 % for verbose queries. However, this is likely and also verifies the second hypothesis in Sect. 4 that XLM would affect the retrieval effectiveness more for verbose queries, because result documents tend to miss more query terms when a query is verbose.

Comparing PRF with short queries, the relative improvements of XML over LM on the former are often larger than those on the latter, which is likely because automatic query expansion using PRF generally improves the verbosity of queries.

Comparing PRF with verbose queries, the relative improvements on the former are often less than those on the latter, although the former often contains more distinct terms (and is thus more verbose) than the latter. One possible reason is that redundant expansion terms may be introduced by PRF. Intuitively, matching a redundant expansion term should not be rewarded as much as matching a basic query term from the original query, suggesting that we may use different δ values for the expansion terms and the basic query terms for PRF, which would be an interesting direction for future work.

We introduce a parameter δ to control the negative query generation component. We plot MAP improvements of XLM over LM against different δ values in Fig. 1. It demonstrates that, for verbose queries, when δ is set to a value around 0.05, XLM works very well across different collections. Therefore, δ can be safely "eliminated" from XLM for verbose queries by setting it to a default value 0.05. Although δ tends to be collection-dependent for short queries and for PRF, setting it conservatively, e.g., 0.02 for short queries and 0.1 for PRF, can often lead to consistent improvement on all collections.

As XLM and LM share one parameter μ, the Dirichlet prior, we are also interested in understanding how this parameter affects the retrieval performance of XLM and LM. So we draw the sensitivity curves of XLM and LM w.r.t. μ in Fig. 2. It shows that XLM is consistently more effective than LM when we vary the value of μ. Moreover, the curve trends of XLM and LM are very similar to each other. In particular, XLM and LM even often share the same optimal setting for μ. These are interesting observations, which suggest that μ and δ do not interact with each other seriously; as a result, we could tune two parameters one by one independently.

We also observe that the optimal μ value for short queries tends to be smaller than that for verbose queries for both LM and XLM, which is consistent to our previous

[5] Following previous work (Zhai and Lafferty 2001a), we use the same query language model to compare XLM and LM so that we can focus on the comparison of scoring functions, i.e., XLM and LM, rather than the feedback techniques. Specifically, the feedback interpolation coefficient $\alpha = 0.8$, the number of feedback documents to 20, the number of terms in feedback model to 50, and the two parameters σ and λ inside the positional relevance model are set to their default values as suggested by Lv and Zhai (2010).

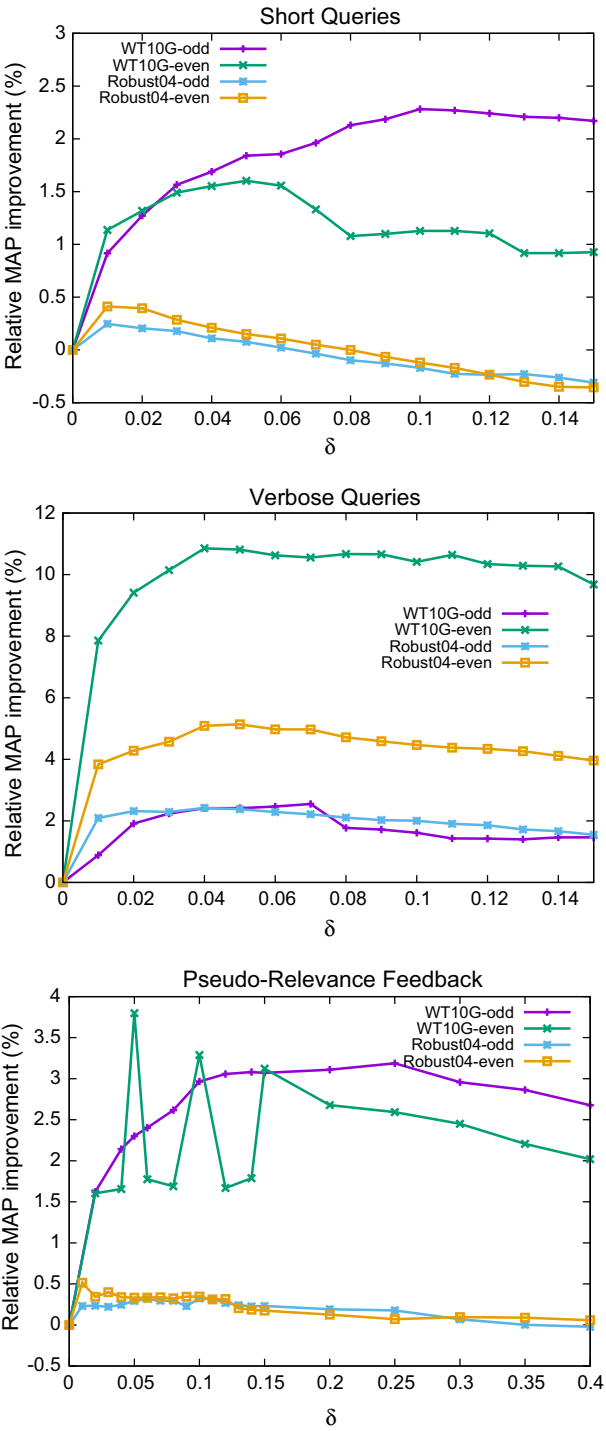

Fig. 1 Performance sensitivity to δ of the proposed method XML for short queries (*top*), verbose queries (*middle*), and query language models estimated using PRF (*bottom*). y-axis shows the relative MAP improvements of XLM over the standard language modeling method. Note that XLM will degenerate to LM when $\delta = 0$. The corresponding settings of parameter μ for each δ value are well tuned. "odd" and "even" in the legend mean that the corresponding curves are based on odd and even-numbered topics respectively

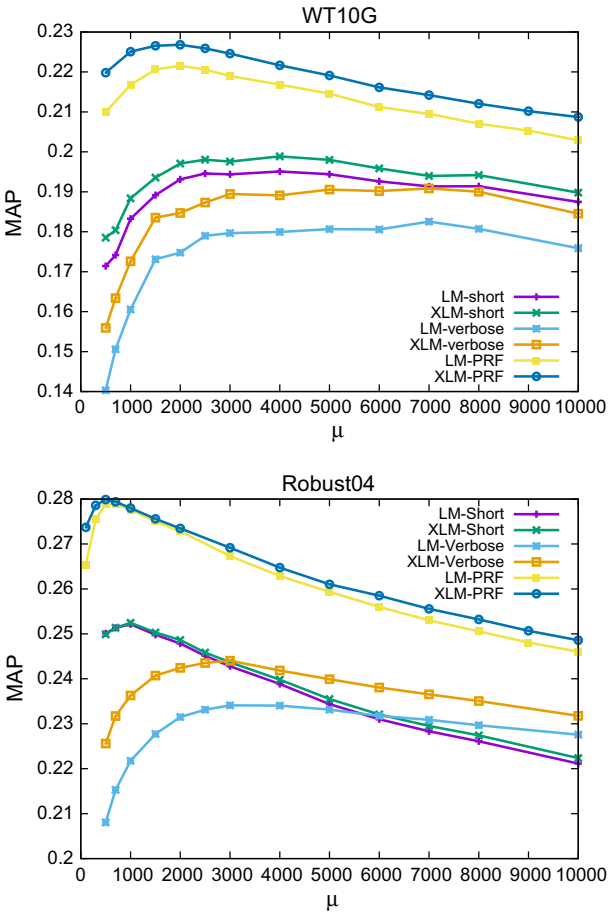

Fig. 2 Performance sensitivity to the Dirichlet prior μ of the standard language model (LM) and the proposed language model with negative query generation (XLM) on WT10G (*top*) and Robust04 (*bottom*). δ is set to 0.05 for short/verbose queries and 0.1 for PRF query language models respectively in XLM, and LM is essentially a special case of XLM where $\delta = 0$. It shows that LM and XLM share similar sensitive curves to μ

observation (Lv and Zhai 2011). Interestingly, we find that the optimal μ for PRF is often smaller than that for short queries, which needs more experiments to understand the reason.

5.3 Summary

Our experiments demonstrate empirically that bringing back the negative query generation component can improve the standard language modeling retrieval models for various types of queries across different collections.

We have derived two effective retrieval functions, a query likelihood model with negative query generation (Formula 21) and a KL-divergence model with negative document language models (Formula 22). Both of them work as efficiently as but more effectively than their corresponding standard retrieval functions, i.e., the standard query likelihood method (Ponte and Croft 1998) and the standard KL-divergence method (Lafferty and Zhai 2001), respectively. There is an extra parameter δ in the derived formulas. However, we observe that δ and the Dirichlet prior μ generally do not interact with each other. Therefore we can tune two parameters independently. The suggested default δ values are 0.05, 0.02 and 0.1 for verbose queries, short queries and PRF, respectively.

6 Conclusions

In this paper, we show that we can improve the standard language modeling retrieval models by bringing back the component of negative query generation (i.e., the probability that a user who dislikes a document would use a query). We argue that ignoring the component of negative query generation in the standard query likelihood retrieval function is questionable, because in reality, a user who dislikes a document would more likely avoid using words in the document when posing a query. We propose an effective approach to estimate document-dependent probabilities of negative query generation based on the principle of maximum entropy, and derive a more complete query likelihood retrieval function that contains the negative query generation component, which essentially scores a document with respect to a query according to the ratio of the probability that a user who likes the document would pose the query to the probability that a user who dislikes the document would pose the query. In addition, we further develop a more general probabilistic distance retrieval method to naturally incorporate query language models, which covers the proposed query likelihood with negative query generation as its special case.

Our work not only bridges the theoretic gap between the standard language modeling retrieval models and the probability ranking principle, but also improves the retrieval effectiveness for various types of queries across different collections, with (almost) no additional computational cost. The proposed retrieval functions can potentially replace the standard query likelihood retrieval method and the standard KL-divergence retrieval method in all retrieval applications.

As a first attempt at bringing back negative query generation, our work opens up an interesting novel research direction in optimizing language models for information retrieval through improving the estimation of negative query generation (or negative document language model). One of the most interesting directions is to study whether setting a term-specific δ can further improve performance. For example, term necessity prediction (Zhao and Callan 2010) and the discovery of key concepts (Bendersky and Croft 2008) could be two possible ways for setting adaptive δ. Another interesting direction is to go beyond document D and seek other resources for estimating a more accurate negative query generation probability.

References

Aly, R., Demeester, T., & Robertson, S. E. (2014). Probabilistic models in ir and their relationships. *Information Retrieval*, *17*(2), 177–201.

Balog, K., Azzopardi, L., & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 43–50). ACM: New York, NY

Bendersky, M., & Croft, W. B. (2008). Discovering key concepts in verbose queries. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08* (pp. 491–498). ACM: New York, NY.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99* (pp. 222–229). ACM: New York, NY.

Collins-Thompson, K., & Callan, J. (2007). Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07* (pp. 303–310). ACM: New York, NY.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02* (pp. 299–306). ACM: New York, NY.

Diaz, F., & Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 154–161). ACM: New York, NY.

Fang, H., & Zhai, C. (2007). Probabilistic models for expert finding. In *Proceedings of the 29th European conference on IR research, ECIR'07* (pp 418–430). Springer: Berlin, Heidelberg.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, *35*, 243–255.

Gao, J., Nie, J. Y., Wu, G., & Cao, G. (2004). Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 170–177). ACM: New York, NY.

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

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

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 178–185). ACM: New York, NY.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Kurland, O., Lee, L., & Domshlak, C. (2005). Better than the real thing? Iterative pseudo-query processing using cluster-based language models. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05* (pp. 19–26). ACM: New York, NY.

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). ACM: New York, NY.

Lafferty, J., & Zhai, C. (2002). Probabilistic relevance models based on document and query generation. In *Language modeling and information retrieval* (pp. 1–10). Kluwer Academic Publishers: Dordrecht.

Lafferty, J. D., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In *Language modeling and information retrieval* (Vol. 13).

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). ACM: New York, NY.

Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02* (pp. 175–182). ACM: New York, NY.

Liu, X., & Croft, W. B. (2002). Passage retrieval based on language models. In *Proceedings of the eleventh international conference on information and knowledge management, CIKM '02* (pp. 375–382). ACM: New York, NY.

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). ACM: New York, NY.

Luk, R. W. (2008). On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval*, *11*(6), 539–561. doi:10.1007/s10791-008-9062-z.

Lv, Y., & Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09* (pp. 299–306). ACM: New York, NY.

Lv, Y., & Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10* (pp. 579–586). ACM: New York, NY.

Lv, Y., & Zhai, C. (2011). Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM international conference on information and knowledge management, CIKM '11* (pp. 7–16). ACM: New York, NY.

Lv, Y., & Zhai, C. (2012). Query likelihood with negative query generation. In *Proceedings of the 21st ACM international conference on information and knowledge management, CIKM '12* (pp. 1799–1803). ACM: New York, NY.

Mei, Q., Fang, H., & Zhai, C. (2007). A study of poisson query generation model for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '07* (pp. 319–326). ACM: New York, NY.

Metzler, D., Lavrenko, V., & Croft, W. B. (2004). Formal multiple-bernoulli models for language modeling. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 540–541). ACM: New York, NY.

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) ACM: New York, NY.

Ogilvie, P., & Callan, J. (2003). Combining document representations for known-item search. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, SIGIR '03* (pp. 143–150). ACM: New York, NY.

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). ACM: New York, NY.

Robertson, S., & Hiemstra, D. (2001). Language models and probability of relevance. In J. Callan, B. W. Croft, & J. Lafferty (Eds.), *Proceedings of the first workshop on language modeling and information retrieval* (pp. 21–25). Pittsburgh, PA: Carnegie Mellon University.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, *33*(4), 294–304.

Robertson, S. E. (2005). On event spaces and probabilistic models in information retrieval. *Information Retrieval*, *8*(2), 319–329.

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

Shen, X., Tan, B., & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '05* (pp. 43–50). ACM: New York, NY.

Si, L., Jin, R., Callan, J., & Ogilvie, P. (2002). A language modeling framework for resource selection and results merging. In *Proceedings of the eleventh international conference on information and knowledge management, CIKM '02* (pp. 391–397). ACM: New York, NY.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on information and knowledge management, CIKM '99* (pp. 316–321). ACM: New York, NY.

Sparck-Jones, K., & Robertson, S. E. (2001). LM vs PM: Where's the relevance? In J. Callan, B. W. Croft, & J. Lafferty (Eds.), *Proceedings of the workshop on language modeling and information retrieval* (pp. 12–15). Pittsburgh, PA: Carnegie Mellon University.

Sparck-Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management*, *36*, 779–808.

Sparck-Jones, K., Robertson, S. E., Hiemstra, D., & Zaragoza, H. (2003). Language modeling and relevance. In B. W. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval, the kluwer international series on information retrieval* (Vol. 13). Dordrecht: Kluwer academic Publishers.

Tan, B., Shen, X., & Zhai, C. (2006). Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06* (pp. 718–723). ACM: New York, NY.

Tao, T., Wang, X., Mei, Q., & Zhai, C. (2006). Language model information retrieval with document expansion. In *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics, HLT-NAACL '06* (pp. 407–414). Association for Computational Linguistics: Stroudsburg, PA.

Tsagkias, M., de Rijke, M., & Weerkamp, W. (2011). Hypergeometric language models for republished article finding. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR '11* (pp. 485–494). ACM: New York, NY.

Wang, X., Fang, H., & Zhai, C. (2008). A study of methods for negative relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08* (pp. 219–226). ACM: New York, NY.

Wei, X., & Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 178–185). ACM: New York, NY.

Xu, J., & Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '99* (pp. 254–261). ACM: New York, NY.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '01* (pp. 105–110). ACM: New York, NY.

Zaragoza, H., Hiemstra, D., & Tipping, M. (2003). Bayesian extension to the language model for ad hoc information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, SIGIR '03* (pp. 4–9). ACM: New York, NY.

Zhai, C. (2008). Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval*, *2*(3), 137–213.

Zhai, C., & Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on information and knowledge management, CIKM '01* (pp. 403–410). ACM: New York, NY.

Zhai, C., & Lafferty, J. (2001b). 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). ACM: New York, NY.

Zhai, C., & Lafferty, J. (2002). Two-stage language models for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02* (pp. 49–56). ACM: New York, NY.

Zhai, C., Cohen, W.W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, SIGIR '03* (pp. 10–17). ACM: New York, NY.

Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02* (pp. 81–88). ACM: New York, NY.

Zhao, L., & Callan, J. (2010). Term necessity prediction. In *Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10* (pp. 259–268). ACM: New York, NY.