Probabilistic IR Models Based on Document and Query Generation

John Lafferty
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lafferty@cs.cmu.edu

Chengxiang Zhai
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
czhai@cs.cmu.edu

Abstract

We present a correspondence between probabilistic retrieval models based on document generation and those based on query generation. We then propose a framework that combines document models and query models using a probabilistic ranking function, casting the retrieval problem in terms of risk minimization.

1. Document Generation vs. Query Generation

We begin by comparing two basic probabilistic retrieval models, one based on document generation and the other based on query generation, showing how each of the two approaches rank documents based on the posterior log-odds ratio of relevance.

In the standard probabilistic approach, the underlying probability model is a generative model of documents from queries (Robertson & Sparck Jones, 1976; Sparck Jones et al., 2000). We suppose that a document is generated from a query using a binary latent variable that indicates whether or not the document is relevant to the query. Following Sparck Jones et al. (2000), we will denote by ℓ (like) and $\bar{\ell}$ (not like) the value of the relevance variable. Thus,

$$p(\mathbf{d}\,|\,\mathbf{q}) = p(\ell\,|\,\mathbf{q})\,p(\mathbf{d}\,|\,\ell,\mathbf{q}) + p(\bar{\ell}\,|\,\mathbf{q})\,p(\mathbf{d}\,|\,\bar{\ell},\mathbf{q})$$

Documents are then ranked according to the posterior log-odds ratio $\log\left(p(\ell\,|\,\mathbf{d},\mathbf{q})/p(\bar{\ell}\,|\,\mathbf{d},\mathbf{q})\right)$. Using Bayes' rule, this ratio can be written as

$$\begin{aligned}
\log\frac{p(\ell\,|\,\mathbf{d},\mathbf{q})}{p(\bar{\ell}\,|\,\mathbf{d},\mathbf{q})} &= \log\frac{p(\mathbf{d},\ell\,|\,\mathbf{q})}{p(\mathbf{d},\bar{\ell}\,|\,\mathbf{q})} \\
&= \log\frac{p(\mathbf{d}\,|\,\ell,\mathbf{q})}{p(\mathbf{d}\,|\,\bar{\ell},\mathbf{q})} + \log\frac{p(\ell\,|\,\mathbf{q})}{p(\bar{\ell}\,|\,\mathbf{q})} \\
&= \log\frac{p(\mathbf{d}\,|\,\ell,\mathbf{q})}{p(\mathbf{d}\,|\,\bar{\ell},\mathbf{q})} + b_{\mathbf{q}}
\end{aligned}$$

where $b_{\mathbf{q}} = \log(p(\ell\,|\,\mathbf{q})/p(\bar{\ell}\,|\,\mathbf{q}))$ is a *bias* constant that depends only on the query, and thus does not affect ranking. One of the motivating considerations for this formal framework is that it satisfies the relevance ranking principle.

Note that if we make the conditional independence assumption $p(\mathbf{d}\,|\,\bar{\ell},\mathbf{q}) = p(\mathbf{d}\,|\,\bar{\ell})$, the log-odds ratio cannot be simplified further, although it can be reexpressed as

$$\log\frac{p(\ell\,|\,\mathbf{d},\mathbf{q})}{p(\bar{\ell}\,|\,\mathbf{d},\mathbf{q})} = \log\frac{p(\mathbf{d}\,|\,\ell,\mathbf{q})}{p(\mathbf{d}\,|\,\bar{\ell})} + b_{\mathbf{q}}$$

The probabilistic approach based on document generation typically then proceeds by assuming document attributes are generated independently, so that the posterior log-odds can be written as a sum of log-odds ratios.

Now, in any interesting inference problem, there will be more than one way in which to apply the chain rule. In what has come to be called "the language modeling approach," the underlying probability model is instead a generative model of queries from documents. While most work in the language modeling approach has not explicitly modeled relevance, let's see what happens when we do. As above, we introduce a binary latent variable that indicates whether or not the document is relevant to the query. The query model $p(\mathbf{q} \,|\, \mathbf{d})$ can then be written as

$$p(\mathbf{q} \,|\, \mathbf{d}) = p(\ell \,|\, \mathbf{d})\, p(\mathbf{q} \,|\, \ell, \mathbf{d}) + p(\bar{\ell} \,|\, \mathbf{d})\, p(\mathbf{q} \,|\, \bar{\ell}, \mathbf{d})$$

Proceeding as before, we rank documents according to the posterior log-odds ratio $\log(p(\ell \,|\, \mathbf{d}, \mathbf{q}) / p(\bar{\ell} \,|\, \mathbf{d}, \mathbf{q}))$. Using Bayes' rule, under the above probabilistic model this ratio can be written as

$$\begin{aligned} \log \frac{p(\ell \,|\, \mathbf{d}, \mathbf{q})}{p(\bar{\ell} \,|\, \mathbf{d}, \mathbf{q})} &= \log \frac{p(\mathbf{q}, \ell \,|\, \mathbf{d})}{p(\mathbf{q}, \bar{\ell} \,|\, \mathbf{d})} \\ &= \log \frac{p(\mathbf{q} \,|\, \ell, \mathbf{d})}{p(\mathbf{q} \,|\, \bar{\ell}, \mathbf{d})} + \log \frac{p(\ell \,|\, \mathbf{d})}{p(\bar{\ell} \,|\, \mathbf{d})} \end{aligned}$$

We now make the following independence assumption: $p(\mathbf{q} \,|\, \bar{\ell}, \mathbf{d}) = p(\mathbf{q} \,|\, \bar{\ell})$; that is, conditioned on the event that the document is not relevant to the query, the query is generated using a model that does not depend on the particular document. Then the log-odds ratio becomes

$$\log \frac{p(\ell \,|\, \mathbf{d}, \mathbf{q})}{p(\bar{\ell} \,|\, \mathbf{d}, \mathbf{q})} = \log p(\mathbf{q} \,|\, \ell, \mathbf{d}) + \log \frac{p(\ell \,|\, \mathbf{d})}{p(\bar{\ell} \,|\, \mathbf{d})} + b'_{\mathbf{q}}$$

where $b'_{\mathbf{q}} = -\log p(\mathbf{q} \,|\, \bar{\ell})$ is a new bias constant that depends only on the query. As before, this formal framework based on query generation satisfies the relevance ranking principle.

At this point we can make some further independence assumptions, assuming that the query terms are generated independently, so that the posterior log-odds can be written as a (penalized) log-likelihood:

$$\log \frac{p(\ell \,|\, \mathbf{d}, \mathbf{q})}{p(\bar{\ell} \,|\, \mathbf{d}, \mathbf{q})} = \sum_i \log p(q_i \,|\, \ell, \mathbf{d}) + \log \frac{p(\ell \,|\, \mathbf{d})}{p(\bar{\ell} \,|\, \mathbf{d})} + b'_{\mathbf{q}}$$

If the prior probability of a document being "liked" is independent of $\mathbf{d}$, then the ranking is determined solely by the log-likelihood term. Researchers have been referring to the distribution $p(\cdot \,|\, \ell, \mathbf{d})$ as a *document language model*, or simply as a *language model*.

In previous discussions of the language modeling approach, an explicit use of a relevance variable has not been made. However, we see from the derivation above that introducing an explicit relevance model is, operationally, of no consequence. This is due to the fact that, unlike in the document generation model, the "irrelevant" language model $p(\mathbf{q} \,|\, \bar{\ell})$ is irrelevant—it only enters into the bias constant, and so can be ignored. We refer to Lavrenko and Croft (2001) for a different introduction of relevance into the language modeling approach.

Assigning interpretations to the language model probabilities appears to have caused great confusion. In particular, it has led to researchers to conclude that the language modeling approach only makes sense if a single document is returned to the user, or that the model becomes inconsistent in the presence of explicit relevance information. However, the derivation just presented shows that in terms of the relevance ranking principle and interpretation of the ranking process, this approach is on equal footing with the traditional probabilistic approach that generates documents from queries. It provides an answer to the question "Where's the relevance?" which has been asked of the language modeling approach.

2. The Importance of Being Reversed

While the derivation given in the previous section puts the language modeling approach and the traditional probabilistic model on equal terms, there are many important differences that result from reversing things to generate the query from the document.

First, in the language modeling approach we have an explicit notion of the importance of a document, represented in the term $\log p(\ell \,|\, \mathbf{d})/(1 - p(\ell \,|\, \mathbf{d}))$. In previous formulations, this role was played by the "document prior" $p(\mathbf{d})$ (Berger & Lafferty, 1999; Miller et al., 1999). While to date the document prior has not been significant for TREC evaluations, for many real applications its use can be expected to be essential. Note, for example, the success of using query-independent scores that assesses the importance of documents based on hyperlink analysis in Web text retrieval (Brin & Page, 1998).

A second important difference is associated with the need for document normalization. In the standard approach, the use of log-odds ratios is essential to account for the fact that different documents have different numbers of terms. Ranking based on document likelihoods $p(\mathbf{d} \mid \ell, \mathbf{q})$ would not be effective because the procedure is inherently biased against long documents. This observation is symptomatic of a larger problem: by making strong independence assumptions we have an incorrect model of relevant documents. In the language modeling approach, things are reversed to generate the input—the query $\mathbf{q}$. As a result, competing documents are scored using the same number probabilities $p(q_i \mid \ell, \mathbf{d})$, and document normalization is not a crucial issue. More generally, incorrect independence assumptions in the model may be mitigated by predicting the input. This advantage of "reverse channel" approaches to statistical natural language processing has been observed in many other applications, most notably statistical machine translation (Brown et al., 1990).

Perhaps the most important consequence of being reversed, however, lies in the fact that by conditioning on the document $\mathbf{d}$, we have a larger foothold for estimating a statistical model. The entire document and, potentially, related documents, can be used to build a language model, and a great deal is known about techniques for estimating such language models from other applications, in particular speech recognition (Bahl et al., 1983). A study of smoothing methods for document language models applied to IR is given in (Zhai & Lafferty, 2001).

3. Combining Query Models and Document Models

We now outline a new formal model that combines document generation models and query generation models. One of the primary motivations is to address what we view as several difficulties with the translation model approach to semantic smoothing in the language modeling framework (Berger & Lafferty, 1999; Berger, 2001). The fundamental problem is that the translation models $t(q \mid w)$ must be estimated from training data. For this purpose, Berger and Lafferty (1999) generate an artificial collection of "synthetic" data. It is unlikely that a sufficiently large collection of relevance judgments will be available to estimate such models on actual user data. In addition, the use of translation models appears to be inefficient, as the model involves a sum over all terms in the document. The translation probabilities are also context-independent, and are therefore unable to directly incorporate word-sense information and context into the language models. Our formal framework is designed to combine language models for the user and specific query, together with document language models. Although the framework is quite general, we will actually only describe it in very special cases.

We view a query as being the output of some probabilistic model associated with a user; similarly, we view a document as being the output of some probabilistic model associated with an author. A query (document) is the result of choosing a model, and then generating the query (document) using that model. The query model could, in principle, encode detailed knowledge about a user's information need and the context in which they make their query. Similarly, the document model could encode complex information about a document and its author, but in its basic form the document model will be a simple multinomial model on terms. To measure the "relevance" of a document to a query, we use a loss function between models, and define relevance in terms of the expected value of this loss, averaged over all models.

More formally, let θ_Q denote the parameters of a query model, and let θ_D denote the parameters of a document model. A user $\mathcal{U}$ generates a query by first selecting θ_Q, according to a distribution $p(\theta_Q \mid \mathcal{U})$. Using this model, a query $\mathbf{q}$ is then generated with probability $p(\mathbf{q} \mid \theta_Q, \mathcal{U})$. Similarly, a (fictitious) author selects a document model θ_D according to a model $p(\theta_D \mid \mathcal{C})$, and then uses this language model to generate a document $\mathbf{d}$.

We assume now that there is a *loss function* $L(\theta_Q, \theta_D)$ between query and document models. If the models θ, θ' are similar, then $L(\theta, \theta')$ should be small. For a given loss function, we define $R(\mathbf{d}; \mathbf{q})$ as the expected loss (posterior mean) under the above probabilistic model. Thus,

$$R(\mathbf{d}; \mathbf{q}) \overset{\text{def}}{=} E\left[L \mid \mathbf{d}, \mathbf{q}\right] = \int_{\Theta_D}\int_{\Theta_Q} L(\theta_Q, \theta_D)\, p(\theta_D \mid \mathbf{d}, \mathcal{C})\, p(\theta_Q \mid \mathbf{q}, \mathcal{U})\, d\theta_Q\, d\theta_D$$

Note that the model selection probabilities $p(\theta_D \mid \mathcal{C})$ and $p(\theta_Q \mid \mathcal{U})$ play a role analogous to priors in Bayesian

$$\begin{tikzcd}[column sep=huge]
\mathcal{U} \arrow[r, "\textsf{\textit{Model selection}}", "p(\theta_Q \,|\, \mathcal{U})"'] & \theta_Q \arrow[r, "\textsf{\textit{Query generation}}", "p(\mathbf{q} \,|\, \theta_Q, \mathcal{U})"'] & \mathbf{q} \\
\mathcal{C} \arrow[r, "\textsf{\textit{Model selection}}", "p(\theta_D \,|\, \mathcal{C})"'] & \theta_D \arrow[r, "\textsf{\textit{Document generation}}", "p(\mathbf{d} \,|\, \theta_D, \mathcal{C})"'] & \mathbf{d}
\end{tikzcd}$$

$$R(\mathbf{d};\mathbf{q}) \overset{\text{def}}{=} E\,[L \,|\, \mathbf{d}, \mathbf{q}]$$

Figure 1. Formal framework for query models and document models. User $\mathcal{U}$ selects a query model θ_Q which is used to generate a query $\mathbf{q}$. Similarly, a fictitious collection author $\mathcal{C}$ creates a document by selecting a document language model with parameters θ_D, and uses this to generate a document $\mathbf{d}$. The risk $R(\mathbf{d};\mathbf{q})$ is defined as the expectation of a loss function $L(\theta_Q, \theta_D)$.

decision theory, and $p(\theta_Q \,|\, \mathbf{q}, \mathcal{U})$ is a posterior probability

$$p(\theta_Q \,|\, \mathbf{q}, \mathcal{U}) \;=\; \frac{p(\mathbf{q} \,|\, \theta_Q, \mathcal{U})\, p(\theta_Q \,|\, \mathcal{U})}{p(\mathbf{q} \,|\, \mathcal{U})}$$

(and similarly for the document model). Because of this connection, and to avoid confusion with the traditional notion of relevance, we will refer to $R(\mathbf{d};\mathbf{q})$ as the *risk* due to answering query $\mathbf{q}$ with document $\mathbf{d}$.

In this framework, a retrieval system is designed to return a document (or documents) $d^\star(\mathbf{q})$ that minimizes the risk

$$d^\star(q) \;=\; \arg\min_{d \in \mathcal{C}} R(\mathbf{d};\mathbf{q})$$

Note that there is not an explicit notion of relevance in the underlying models; rather relevance enters only in the particular loss function used. Moreover, note that the risk $R(\mathbf{d};\mathbf{q})$ for a given document is not assumed to be independent of the risk of other documents in the collection; indeed, the prior $p(\theta_D \,|\, \mathcal{C})$ will in general take the entire document collection into account. Of course, the true user model $p(\theta_Q \,|\, \mathcal{U})$ and author model $p(\theta_D \,|\, \mathcal{C})$ are not known. Typically we have only a small collection of relevance judgments $(\mathbf{q}_i, \mathbf{d}_i)$ at our disposal to help infer these models.

As a special case of the risk minimization framework, assume that θ_Q and θ_D are the parameters of unigram language models. Thus, $p(\cdot \,|\, \theta)$ is a distribution over a fixed word vocabulary. We choose as the loss function the Kullback-Leibler divergence:

$$L(\theta_Q, \theta_D) \;=\; \sum_w p(w \,|\, \theta_Q) \log \frac{p(w \,|\, \theta_Q)}{p(w \,|\, \theta_D)}$$

In this case, the risk R is the expected KL divergence, averaged over all document and query models, weighted by the posterior probabilities.

Now, for a word list $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$, let $\widetilde{\theta}_{\mathbf{w}}$ be the *empirical distribution* of the list. That is, $\widetilde{\theta}_{\mathbf{w}}$ is the language model given by

$$p(w \,|\, \widetilde{\theta}_{\mathbf{w}}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \delta(w, w_i)$$

Assume now that the query posterior probability is highly concentrated on a single point $\widehat{\theta}_{\mathbf{q}}$, and similarly that the document posterior probability is concentrated on a single point $\widehat{\theta}_{\mathbf{d}}$. Then the risk is approximated by a weighted KL divergence. That is (dropping the dependence on $\mathcal{U}$ and $\mathcal{C}$ in the notation),

$$R(\mathbf{d};\mathbf{q}) \;\approx\; -p(\widehat{\theta}_{\mathbf{q}} \,|\, \mathbf{q})\, p(\widehat{\theta}_{\mathbf{d}} \,|\, \mathbf{d}) \sum_w p(w \,|\, \widehat{\theta}_{\mathbf{q}}) \log p(w \,|\, \widehat{\theta}_{\mathbf{d}}) + c_{\mathbf{q}}$$

where $c_{\mathbf{q}}$ is a constant that doesn't depend on the document, and so doesn't affect the retrieval performance. Note that the factor $p(\widehat{\theta}_{\mathbf{d}} \,|\, \mathbf{d})$ includes prior information about the document, and in general must be included when comparing the risk for different documents. This is critical when incorporating query-independent link analysis, or other extrinsic knowledge about a document.

If now $\widehat{\theta}_{\mathbf{q}}$ is just the empirical distribution of the query, $\widetilde{\theta}_{\mathbf{q}}$, and the priors are uniform, then we obtain

$$R(\mathbf{d};\mathbf{q}) \quad \approx \quad -\frac{1}{m}\sum_{i=1}^{m} \log p\,(q_i \,|\, \widehat{\theta}_{\mathbf{d}}) + c_{\mathbf{q}}$$

This is precisely the log-likelihood criterion that has been used in all work on the language modeling approach to date. Several other well-known IR frameworks can be derived as special cases of risk minimization.

In (Lafferty & Zhai, 2001) a query expansion method using language models is derived and applied in the risk minimization framework, and tested empirically on TREC benchmarks with promising results.

References

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *5*, 179–190.

Berger, A. (2001). *Statistical machine learning for information retrieval.* Doctoral dissertation, Carnegie Mellon University.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)* (pp. 222–229).

Brin, S., & Page, L. (1998). Anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference.*

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*, 79–85.

Lafferty, J., & Zhai, C. (2001). Risk minimization and Markov chain query expansion in the language modeling approach to information retrieval. *24th ACM SIGIR Conference on Research and Development in Information Retrieval.* To appear.

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. *24th ACM SIGIR Conference on Research and Development in Information Retrieval.* To appear.

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)* (pp. 214–221).

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

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments, Part 1. *Information Processing and Management*, *36*, 779–808.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *24th ACM SIGIR Conference on Research and Development in Information Retrieval.* To appear.

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.