Probabilistic Relevance 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

June 17, 2002

Abstract

We give a unified account of the probabilistic semantics underlying the language modeling approach and the traditional probabilistic model for information retrieval, showing that the two approaches can be viewed as being equivalent probabilistically, since they are based on different factorizations of the same generative relevance model. We also discuss how the two approaches lead to different retrieval frameworks in practice, since they involve component models that are estimated quite differently.

1. Introduction

In the classical probabilistic approach to information retrieval (Robertson & Sparck Jones, 1976), two models are estimated for each query, one modeling relevant documents, the other modeling non-relevant documents.[1] Documents are then ranked according to the posterior probability of relevance. When the document attributes are independent under these relevance models, this is simply the naive Bayes model for classification, and has met with considerable empirical success.

In the "language modeling approach" to information retrieval (Ponte & Croft, 1998), a language model is estimated for each document, and the operational procedure for ranking is to order documents by the probability assigned to the input query text according to each document's model. This approach has also enjoyed recent empirical success. However, the underlying semantics of the language model has been unclear, as it appears to ignore the important notion of relevance.

In this paper we give a simple, unified account of both approaches, in which it is shown that an implicit relevance model underlies the language modeling approach. Our derivation shows that the two approaches are in fact equivalent probabilistically, since they are based on different parameterizations of the same joint likelihood. However, as we discuss below, the two approaches

[1] Sparck Jones et al. (2000) refer to this as *the* probabilistic approach to retrieval.

are not equivalent from a statistical point of view, since the component models are estimated quite differently.

Our derivation is elementary, and shows that in terms of their underlying probabilistic semantics, the language modeling approach and the traditional probabilistic model are, so to speak, two sides of the same coin. Thus, we provide a simple answer to the question "Where's the relevance?" that has been recently asked of the language modeling approach.

2. Generative Relevance Models

2.1. The Basic Question

In our treatment of the probabilistic semantics of relevance models, we follow the presentation of Sparck Jones et al. (2000), with some minor changes in notation. Thus, the "Basic Question" we are interested in is the following:

> What is the probability that *this* document is relevant to *this* query?

To treat the Basic Question in a probabilistic framework, we introduce random variables D and Q to denote a document and query, respectively. In addition, we introduce a binary random variable R to denote relevance.[2] This random variable takes on two values, which we denote as r ("relevant") and $\overline{r}$ ("not relevant"). Here our notation deviates from that of Sparck Jones et al. (2000), who use L ("liked") and $\overline{L}$ ("not liked") instead of r and $\overline{r}$. We thus adopt the standard notation that denotes random variables using upper case letters and values of random variables using lower case letters. In probabilistic terms, the Basic Question is then equivalent to estimating the probability of relevance

$$p\,(R = r \mid D, Q) \;\; = \;\; 1 - p\,(R = \overline{r} \mid D, Q)\,. \tag{2.1}$$

The justification for using this probability as the basis for ranking comes from the Probability Ranking Principle (Robertson, 1977).

Now, in adopting a *generative* relevance model, the probability of relevance $p\,(r \mid D, Q)$ is not estimated directly. Rather, it is estimated indirectly by invoking Bayes' rule:

$$p\,(R = r \mid D, Q) \;\; = \;\; \frac{p\,(D, Q \mid R = r)\, p\,(R = r)}{p\,(D, Q)}\,. \tag{2.2}$$

Equivalently, we may use the following log-odds ratio to rank documents:

$$\log \frac{p\,(r \mid D, Q)}{p\,(\overline{r} \mid D, Q)} \;\; = \;\; \log \frac{p\,(D, Q \mid r)\, p\,(r)}{p\,(D, Q \mid \overline{r})\, p\,(\overline{r})}\,. \tag{2.3}$$

As we describe next, two statistically different but probabilistically equivalent generative relevance models result from applying the chain rule in different ways to factor the conditional probability $p\,(D, Q \mid R)$.

[2] Sparck Jones et al. (2000) use R to denote the number of relevant documents.

2.2 The Robertson-Sparck Jones Model

In the Robertson-Sparck Jones approach (Sparck Jones et al., 2000), the probability $p(D, Q \,|\, R)$ is factored as $p(D, Q \,|\, R) = p(Q \,|\, R)\, p(D \,|\, Q, R)$, leading to the following log-odds ratios:

$$
\begin{aligned}
\log \frac{p(r \,|\, D, Q)}{p(\overline{r} \,|\, D, Q)} &= \log \frac{p(D, Q \,|\, r)\, p(r)}{p(D, Q \,|\, \overline{r})\, p(\overline{r})} && (2.4) \\
&= \log \frac{p(D \,|\, Q, r)\, p(Q \,|\, r)\, p(r)}{p(D \,|\, Q, \overline{r})\, p(Q \,|\, \overline{r})\, p(\overline{r})} && (2.5) \\
&= \log \frac{p(D \,|\, Q, r)\, p(r \,|\, Q)}{p(D \,|\, Q, \overline{r})\, p(\overline{r} \,|\, Q)} && (2.6) \\
&= \log \frac{p(D \,|\, Q, r)}{p(D \,|\, Q, \overline{r})} + \log \frac{p(r \,|\, Q)}{p(\overline{r} \,|\, Q)} && (2.7) \\
&\stackrel{\text{rank}}{=} \log \frac{p(D \,|\, Q, r)}{p(D \,|\, Q, \overline{r})}\,. && (2.8)
\end{aligned}
$$

Since the term $\log(p(r \,|\, Q)/p(\overline{r} \,|\, Q))$ is independent of D, it can be thought of as a constant bias and can be safely ignored for the purpose of ranking documents; this equivalence is denoted by the symbol $\stackrel{\text{rank}}{=}$.

Equation (2.7) is precisely the basic ranking formula (1) in (Sparck Jones et al., 2000), although the conditioning on the query Q is implicit there.

In its usual instantiation, the models $p(D \,|\, Q, r)$ and $p(D \,|\, Q, r)$ are estimated by assuming that the document is made up of a collection of attributes $D = (A_1, \ldots, A_n)$, such as words, and that these attributes are independent given R and Q:

$$
\begin{aligned}
p(D \,|\, Q, r) &= \prod_{i=1}^{n} p(A_i \,|\, Q, r) && (2.9) \\
p(D \,|\, Q, \overline{r}) &= \prod_{i=1}^{n} p(A_i \,|\, Q, \overline{r})\,. && (2.10)
\end{aligned}
$$

For a fixed query Q, this is simply the naive Bayes model for classifying documents into the two classes r and $\overline{r}$.

2.3 The Language Modeling Approach

Suppose that we now factor the probability $p(D, Q \,|\, R)$ as $p(D, Q \,|\, R) = p(D \,|\, R)\, p(Q \,|\, D, R)$. It is important to note that from a purely probabilistic perspective, nothing has changed; this is simply a different decomposition of the same joint likelihood. Using this factorization, we are led to consider the log-odds ratio in the following *equivalent* form:

$$
\log \frac{p(r \,|\, Q, D)}{p(\overline{r} \,|\, Q, D)} = \log \frac{p(D, Q \,|\, r)\, p(r)}{p(D, Q \,|\, \overline{r})\, p(\overline{r})} \qquad (2.11)
$$

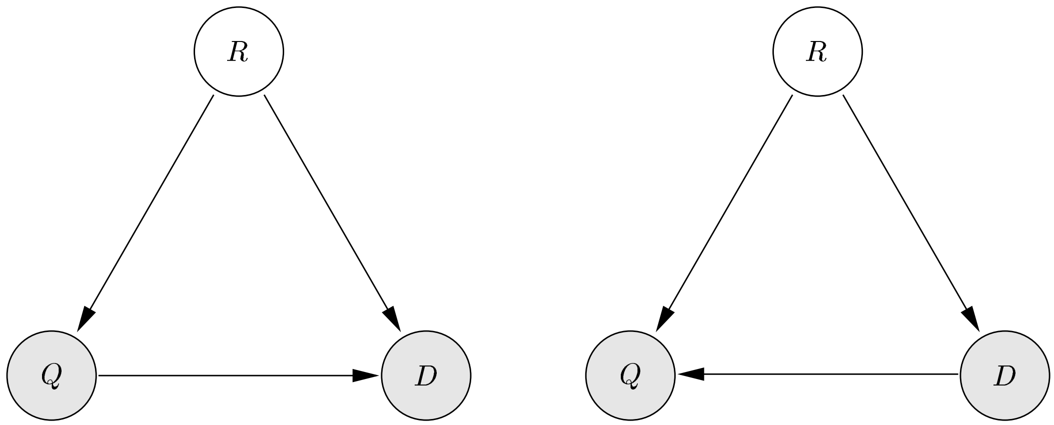

Figure 1: Graphical representations of the two factorizations of the joint document-query probability. The factorization $p(D, Q \mid R) = p(D \mid Q, R)\, p(Q \mid R)$ (left) results in the Robertson-Sparck Jones model, while the factorization $p(D, Q \mid R) = p(Q \mid D, R)\, p(D \mid R)$ (right) leads to the language modeling approach. Following convention, the document and query nodes are shaded to indicate that they are observed ("*this* document and *this* query").

$$= \log \frac{p(Q \mid D, r)\, p(D \mid r)\, p(r)}{p(Q \mid D, \overline{r})\, p(D \mid \overline{r})\, p(\overline{r})} \tag{2.12}$$

$$= \log \frac{p(Q \mid D, r)\, p(r \mid D)}{p(Q \mid D, \overline{r})\, p(\overline{r} \mid D)} \tag{2.13}$$

$$= \log \frac{p(Q \mid D, r)}{p(Q \mid D, \overline{r})} + \log \frac{p(r \mid D)}{p(\overline{r} \mid D)}\,. \tag{2.14}$$

The bias term $\log(p(r \mid D)/p(\overline{r} \mid D))$ is now dependent on D, but independent of the query Q, and must, in general, be considered as an integral part of the ranking process. At this point, we have a ranking formula based on generating queries from documents that is equivalent to the ranking formula (2.7) based on generating documents from queries.

Suppose that we now make the assumption that conditioned on the event $R = \overline{r}$, the document D is independent of the query Q; that is:

Assumption 1: $p(D, Q \mid R = \overline{r}) \;=\; p(D \mid R = \overline{r})\, p(Q \mid R = \overline{r})$

Under this assumption the log-odds ratio becomes

$$\log \frac{p(r \mid Q, D)}{p(\overline{r} \mid Q, D)} = \log \frac{p(Q \mid D, r)}{p(Q \mid \overline{r})} + \log \frac{p(r \mid D)}{p(\overline{r} \mid D)} \tag{2.15}$$

$$\stackrel{\text{rank}}{=} \log p(Q \mid D, r) + \log \frac{p(r \mid D)}{p(\overline{r} \mid D)}\,. \tag{2.16}$$

This ranking formula has two components, a term involving the query likelihood $p(Q \mid D, r)$, and a bias term that involves the prior probability of relevance for the document, $p(r \mid D)$. Researchers

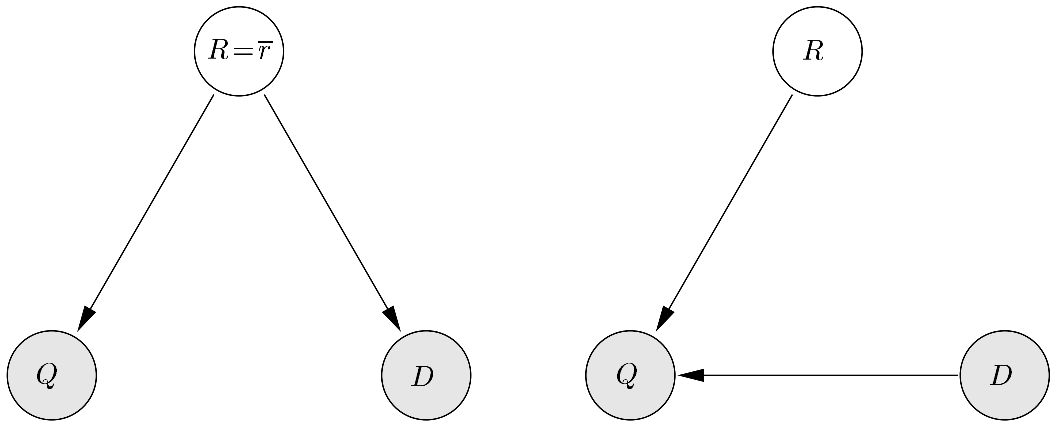

Figure 2: Graphical representations of the document-query distribution under Assumption 1 (left), and Assumption 2 (right).

have been referring to the distribution $p(\cdot \mid D, r)$ as a "document language model," or simply as a "language model," which is actually a model of the *queries* to which D would be judged as relevant. Although the Robertson-Sparck Jones approach also makes use of language models, the terminology "language modeling approach" is appropriate for this way of decomposing the document-query probability since many language models are at play, at least one for each document in the database.

If we now make the additional assumption that D and R are independent, the bias term no longer depends on D. That is, under Assumption 1 and

$$\text{Assumption 2:} \quad p(D, R) \;=\; p(D)\, p(R)$$

the log-odds ratio becomes

$$\log \frac{p(r \mid Q, D)}{p(\overline{r} \mid Q, D)} \;\stackrel{\text{rank}}{=}\; \log p(Q \mid D, r) + \log \frac{p(r)}{p(\overline{r})} \tag{2.17}$$

$$\stackrel{\text{rank}}{=}\; \log p(Q \mid D, r)\,. \tag{2.18}$$

Thus, the ranking of documents is based solely on the probability of the query given the document, under that event that the document is relevant to the query: $p(Q \mid D, r)$. The above assumptions are shown graphically in Figure 2.

Equations (2.16) and (2.18) are the basic ranking formulas for the language modeling approach as explored in (Berger & Lafferty, 1999; Miller et al., 1999) and (Ponte & Croft, 1998) respectively.

As in the Robertson-Sparck Jones model, it is expedient to decompose the query into attributes $Q = (A_1, \ldots, A_m)$, typically just the query terms, and to assume that the attributes are independent given R and the document. Thus, under this independence assumption and Assumption 1, the posterior log-odds becomes

$$\log \frac{p(r \mid D, Q)}{p(\overline{r} \mid D, Q)} \;\stackrel{\text{rank}}{=}\; \sum_{i=1}^{m} \log p(A_i \mid D, r) + \log \frac{p(r \mid D)}{p(\overline{r} \mid D)}\,. \tag{2.19}$$

3. Discussion

The above derivation shows that the language modeling approach and the traditional probabilistic model can be interpreted within the same probabilistic framework based on a generative relevance model. In this view, the two approaches are simply two sides of the same coin—at the probabilistic level, before independence assumptions are made and without any specification of how the models are actually estimated, they are equivalent.

In previous discussions of the language modeling approach in the literature, an explicit use of a relevance variable has not been made. However, under Assumption 1, which states that $p(D, Q \mid R = \overline{r}) = p(D \mid R = \overline{r})\, p(Q \mid R = \overline{r})$, it is seen that introducing an explicit relevance model is operationally of no consequence—the "irrelevant" language model $p(Q \mid \overline{r})$ is irrelevant; it only enters into the bias term, and so can be ignored for ranking. Note that under the same Assumption 1, the log-odds ratios in the Robertson-Sparck Jones approach still involve models for both relevant and non-relevant documents, but now the model for non-relevant documents is simply independent of the query:

$$\log \frac{p(r \mid D, Q)}{p(\overline{r} \mid D, Q)} \stackrel{\text{rank}}{=} \log \frac{p(D \mid Q, r)}{p(D \mid Q, \overline{r})} = \log \frac{p(D \mid Q, r)}{p(D \mid \overline{r})}\,. \tag{3.1}$$

During discussions at the Language Modeling and Information Retrieval Workshop (Callan et al., 2001), it became clear that descriptions of the language modeling approach in terms of generative models of queries have caused significant confusion. In particular, such descriptions have led some researchers to claim that the language modeling approach only makes sense if there is exactly one relevant document for each query, and that the model becomes inconsistent in the presence of explicit relevance information from a user. However, as the above presentation makes clear, the underlying probabilistic semantics is the same as for the standard probabilistic model.

While the derivation presented in the previous section clarifies the formalism behind the language modeling approach, it is only formalism. The genius of a statistical approach often lies in the estimation details. Several important differences result from reversing things to generate the query from the document, which make this approach attractive from an estimation perspective.

Perhaps the primary importance of being reversed lies in the fact that by conditioning on the document 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. Intuitively, it is easier to estimate a model for "relevant queries" based on a document than to estimate a model for relevant documents based on a query. Indeed, the Robertson-Sparck Jones model has encountered difficulties in estimating $p(A_i \mid Q, r)$ and $p(A_i \mid Q, \overline{r})$ when no explicit relevance information is available. Typically, $p(A_i \mid Q, r)$ is set to a constant and $p(A_i \mid Q, \overline{r})$ is estimated under the assumption that the entire collection is comprised of non-relevant documents (Croft & Harper, 1979; Robertson & Walker, 1997)—essentially the same as Assumption 1. Recently, a better approach to estimating $p(A_i \mid Q, r)$ is proposed in (Lavrenko & Croft, 2001), in which the query is formally treated as an observation from the model of generating relevant documents, and a set of empirical document language models are exploited to smooth the estimate of $p(A_i \mid Q, r)$. This work can be considered as an example of using language models in the classical probabilistic model.

Another potential advantage lies in the fact that the language modeling approach includes an explicit notion of the importance of a document, represented in the term $\log p(r \mid D)/(1 - p(r \mid D))$, which can be estimated separately. In previous formulations, this role was played by the "document prior" (Berger & Lafferty, 1999; Miller et al., 1999). While to date the document prior has not been significant for TREC-style evaluations, for many real applications its use can be expected to be important. In particular, query-independent scores to assess the importance of documents based on hyperlink analysis have proven to be useful in web search (Brin & Page, 1998).

An additional difference between the two approaches lies in 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 $p(D \mid r, Q)$ 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 Q. As a result, competing documents are scored using the same number probabilities $p(A_i \mid D, r)$, 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, notably statistical machine translation (Brown et al., 1990).

Having mentioned some of the advantages of query-generation models, we should add that the Robertson-Sparck Jones model, being based on document-generation, has the advantage of being able to naturally improve the estimation of the component probabilistic models by exploiting explicit relevance information. This is because the relevance judgments from a user provide direct training data for estimating $p(A_i \mid Q, r)$ and $p(A_i \mid Q, \overline{r})$, which can then be applied to *new* documents. The same relevance judgments can also provide direct training data for improving the estimate of $p(A_i \mid D, r)$ in the language modeling approach, but only for the relevant documents that are given judgements. Thus, the directly improved models can *not* be expected to improve our ranking of other unjudged documents. However, such improved models can potentially be beneficial for new queries, a feature that does not apply to document-generation models.

4. Historical Notes

Interestingly, the very first probabilistic model for information retrieval, namely the Probabilistic Indexing model of Maron and Kuhns (Maron & Kuhns, 1960) is, in fact, based on the idea of "query-generation." Conceptually, the model intends to infer the probability that a document is relevant to a query based on the probability that a user who likes the document would have used this query. However, the formal derivation given in (Maron & Kuhns, 1960) appears to be restricted to queries with only a single term. As a result, the "query-generation" model $p(w \mid D, r)$ essentially provides a probability for each indexing word, and can be used as a basis for assigning indexing terms to the document. Thus, it is referred to as a probabilistic *indexing* model. Possibly due to its restriction to single-term queries and the difficulty of parameter estimation, this model has never been as popular as the Robertson-Sparck Jones model. However conceptually, they can be considered as representing the two major types of classical probabilistic models.

There were some early efforts to unify these two classical probabilistic models (e.g., (Robertson et al., 1982)), but the unification was not completely successful. The difficulty encountered in (Robertson et al., 1982) has to do with using a more restricted event space, namely a space given by the cross product of documents and queries, without the relevance variable. No doubt, this early work already recognizes the symmetry between queries and documents. See (Robertson, 1994) further discussion of this symmetry and objections to it.

The possibility of both document-generation and query-generation decompositions of the same probability of relevance was also recognized at least a decade ago. Indeed, the two different decompositions were already used in (Fuhr, 1992) to derive, respectively, the Robertson-Sparck Jones model and the Binary Independence Indexing (BII) model, which is a variant of the original Maron and Kuhns model that allows multi-word queries.

References

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.

Callan, J., Croft, B., & Lafferty, J. (Eds.). (2001). *Proceedings of the Workshop on Language Modeling and Information Retrieval.* Carnegie Mellon University.

Croft, W. B., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, *35*, 285–295.

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

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

Maron, M., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, *7*, 216–244.

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).

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. *Proceedings of the 21st International Conference on Research and Development in Information Retrieval (SIGIR'98)* (pp. 275–281).

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

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

Robertson, S., & Walker, S. (1997). On relevance weights with little relevance information. *Proceedings of SIGIR'97* (pp. 16–24).

Robertson, S. E. (1994). Query-document symmetry and dual models. *Journal of Documentation*, *50*, 233–238.

Robertson, S. E., Maron, M. E., & Cooper, W. S. (1982). Probability of relevance: a unification of two competing models for information retrieval. *Information Technology - Research and Development*, *1*, 1–21.

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.