A Risk Minimization Framework for Information Retrieval

ChengXiang Zhai [a] John Lafferty [b]

[a] *Department of Computer Science*
University of Illinois at Urbana-Champaign
[b] *School of Computer Science*
Carnegie Mellon University

Abstract

This paper presents a probabilistic information retrieval framework in which the retrieval problem is formally treated as a statistical decision problem. In this framework, queries and documents are modeled using statistical language models, user preferences are modeled through loss functions, and retrieval is cast as a risk minimization problem. We discuss how this framework can unify existing retrieval models and accommodate systematic development of new retrieval models. As an example of using the framework to model non-traditional retrieval problems, we derive retrieval models for subtopic retrieval, which is concerned with retrieving documents to cover many different subtopics of a general query topic. These new models differ from traditional retrieval models in that they relax the traditional assumption of independent relevance of documents.

Key words:
Retrieval models, risk minimization, statistical language models, Bayesian decision theory

1 Introduction

Over the course of decades of research in information retrieval, many different information retrieval models have been proposed and studied. While significant progress has been made, no single retrieval model has proven to be most effective, and several major challenges remain. For example, theoretical guidelines and formal principles have rarely led to good performance directly; instead, a theoretically well defined formula often needs to be heuristically modified in order to perform well empirically. It is thus a significant scientific challenge to develop principled retrieval approaches that also perform well empirically. In addition, most retrieval models have been developed based on the assumption of independent relevance –

the relevance value of a document is assumed to be independent of that of other documents, including those already viewed by a user. Clearly, this assumption does not hold in real applications. A major challenge is to develop models that can relax such an assumption.

In this paper, we present a probabilistic information retrieval framework that begins to address these challenges. The basic idea of the framework is to formally treat the task of information retrieval as a statistical decision problem. Specifically, given a collection of documents, a query, and any other information that we know about the user, a retrieval system needs to choose a subset of documents and present them in an appropriate way. For example, ranking all the documents according to a query, as is done in a typical retrieval system, can be regarded as a decision problem where the decision involves choosing the best ranking. We formalize this view of retrieval using Bayesian decision theory. In particular, we treat a query and a document as observations from a probabilistic model, called a statistical language model, and encode retrieval preferences with a loss function defined on the language models and a retrieval action. According to Bayesian decision theory, the optimal retrieval action (e.g., the optimal ranking in the case when the decision involves choosing a ranking) is the one that minimizes the Bayes risk, which is the expected loss associated with the chosen action given the observed query and documents.

This framework unifies several existing retrieval models, including the recently proposed language modeling approach, within a single probabilistic framework, and provides guidance on how one can further improve a retrieval model and systematically explore new approaches to information retrieval. Several new retrieval models derived using the risk minimization framework have been shown to be quite effective empirically.

In addition to its generality, this risk minimization framework has several potential advantages over a traditional formulation of the information retrieval problem. First, it systematically incorporates statistical language models as components in a retrieval framework. Statistical language models provide a principled way to model text documents and queries, making it possible to set retrieval parameters through statistical inference and estimation methods. Second, the risk minimization framework makes it possible to systematically and formally study optimal retrieval strategies. For example, through making different assumptions about the loss function for ranking we can derive an optimal ranking principle, which is similar to the probability ranking principle, but which addresses several limitations of this standard principle. Finally, the risk minimization framework extends the traditional notion of independent, topical relevance. For example, it is possible to formalize retrieval models for a non-traditional retrieval task where the goal is to retrieval as many different subtopics of a general topic as possible.

The rest of the paper is organized as follows. In Section 2, we briefly review existing retrieval models and discuss how the risk minimization framework is related to

them. In Section 3, we present the basic idea and setup of the risk minimization framework. In Section 4.1 and Section 4.2, we derive several special cases of the framework, demonstrate how it can cover existing retrieval models and also how it can facilitate development of new retrieval models, including those appropriate for the non-traditional subtopic retrieval task, as discussed in detail in Section 5. Finally, we summarize the contributions of the paper in Section 6 and Section 7.

2 Existing Retrieval Models

Through years of research, many different retrieval models have been proposed, studied, and tested. Their mathematical basis spans a large spectrum, including algebra, logic, set theory, and probability and statistics. Although it is impractical to provide a complete survey of all the existing retrieval models in this paper, we can roughly classify the existing models into three broad categories, depending on how they define and measure relevance. In one category, relevance is assumed to be correlated with the similarity between a query and a document. In another category, a binary random variable is used to model relevance and probabilistic models are used to estimate the value of this relevance variable. In the third category, the uncertainty of relevance is modeled by the uncertainty in inferring queries from documents or vice versa. In order to place the risk minimization framework in context, we discuss each of these three categories below.

2.1 Similarity-based Models

In a similarity-based retrieval model (Dominich, 2000, 2001), it is assumed that the relevance status of a document with respect to a query is correlated with the similarity between the query and the document at some level of representation; the more similar to a query a document is, the more relevant the document is assumed to be. In practice, we can use any similarity measure that preserves such correlation to generate a relevance status value (RSV) for each document and rank documents accordingly.

The vector space model is the most well known model of this type (Salton et al., 1975a; Salton and McGill, 1983; Salton, 1989), in which a document and a query are represented as two term vectors in a high-dimensional term space and each term is assigned a weight that reflects its "importance" to the document or the query. Given a query, the relevance status value of a document is given by the similarity between the query vector and document vector as measured by some vector similarity measure, such as the cosine of the angle formed by the two vectors.

The vector space model naturally decomposes a retrieval model into three com-

ponents: (1) a term vector representation of a query; (2) a term vector representation of a document; (3) a similarity/distance measure between a document vector and a query vector. However, the "synchronization" among the three components is generally unspecified; in particular, the similarity measure does not dictate the representation of a document or query. Thus, the vector space model is actually a general retrieval *framework*, in which the representation of query and documents as well as the similarity measure are all, in principle, arbitrary.

The flexibility of the vector space approach makes it easy to incorporate different indexing models. For example, the 2-Poisson probabilistic indexing model can be used to select indexing terms or assign term weights (Harter, 1975; Bookstein and Swanson, 1975). Latent semantic indexing can be applied to reduce the dimension of the term space and to capture the semantic "closeness" among terms, in an effort to improve the representation of the documents and query (Deerwester et al., 1990). A document can also be represented by a multinomial distribution over the terms, as in the distribution model of indexing proposed in (Wong and Yao, 1989).

The main criticism of the vector space model is that it provides no formal framework for the representation, making the study of representation inherently separate from the estimation of relevance. The separation of the relevance function from the weighting of terms has the advantage of being flexible, but the disadvantage of making difficult the study of the interaction between representation and relevance measurement. The optimality of a similarity/relevance function is highly dependent on the actual representation (i.e., term weights) of the query and the document. As a result, the study of representation in the vector space model has been largely heuristic. The two central problems in document and query representation are the extraction of indexing terms, or other units, and the weighting of the indexing terms. The choice of different indexing units has been extensively studied, but no significant improvement has been achieved over the simplest word-based indexing (Lewis, 1992), although recent evaluation has shown more promising improvement through the use of linguistic phrases (Evans and Zhai, 1996; Strzalkowski, 1997; Zhai, 1997). Many heuristics have also been proposed to improve term weighting, but again, no weighting method has been found to be significantly better than the heuristic TF-IDF term weighting (Salton and Buckley, 1988). To address the variance in the length of documents, an effective weighting formula also needs to incorporate document length heuristically (Singhal et al., 1996). Salton et al. introduced the idea of the *discrimination value* of an indexing term (Salton et al., 1975b), which is the increase or decrease in the mean inter-document distance caused by adding the indexing term to the term space for text representation. Salton et al. found that the middle frequency terms have higher discrimination value. Given a similarity measure, the discrimination value provides a principled way of selecting terms for indexing. However, there are still two deficiencies. First, the framework is not modeling relevance, but rather relies on a fixed similarity measure. Second, it is only helpful for selecting indexing terms, but not for the weighting of terms. Other criticisms about the vector-space model can be found in

(Bollmann-Sdorra and Raghavan, 1993; Dominich, 2002).

As seen below, the risk minimization framework suggests a new formal similarity-based retrieval model in which the representation of query and documents is associated with statistical language models. The use of statistical language models makes it possible to replace the traditional ad hoc tuning of parameters with statistical estimation of parameters.

2.2 Probabilistic Relevance Models

In a probabilistic relevance model, one is interested in the question "What is the probability that *this* document is relevant to *this* query?" (Sparck Jones et al., 2000). Given a query, a document is assumed to be either relevant or non-relevant, but the system relies on a probabilistic model to infer this value.

Formally, let random variables D and Q denote a document and query, respectively. Let R be a binary random variable that indicates whether D is relevant to Q or not. It takes two values which we denote as r ("relevant") and $\overline{r}$ ("not relevant"). The task is to estimate the probability of relevance, i.e., $p(R = r \mid D, Q)$. Depending on how this probability is modeled and estimated, there are several special cases of this general probabilistic relevance model.

First, $p(R = r \mid D, Q)$ can be estimated directly using a discriminative (regression) model. Essentially, the relevance variable R is assumed to be dependent on "features" that characterize how well D matches Q. Such a regression model was first introduced, with some success, by Fox (1983), where features such as term frequency, authorship, and co-citation were combined using linear regression. Fuhr and Buckley (1991) used polynomial regression to approximate relevance. Gey used logistic regression involving information such as query term frequency, document term frequency, IDF, and relative term frequency in the whole collection, and this model shows promising performance in three small testing collections (Gey, 1994). Regression models provide a well studied framework in which to explore the use of heuristic features. One important advantage of regression models is their ability to learn from all the past relevance judgments, in the sense that the parameters of a model can be estimated based on all the relevance judgments, including the judgments for different queries or documents. However, a large amount of data and empirical experimentation may be needed in order to find a set of good features. The regression framework thus provides only limited guidance for extending a retrieval model.

Alternatively, $p(R = r \mid D, Q)$ can be estimated indirectly using a generative model, and documents can be ranked according to the following log-odds ratio:

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

There are two different ways to factor the conditional probability $p(D, Q \mid R)$, corresponding to *document generation* and *query generation* (Lafferty and Zhai, 2003). Most classic probabilistic retrieval models (Robertson and Sparck Jones, 1976; van Rijsbergen, 1979; Robertson et al., 1981; Fuhr, 1992) are based on document generation (i.e., $p(D, Q \mid R) = p(D \mid Q, R)p(Q \mid R)$). The Binary Independence Retrieval (BIR) model (Robertson and Sparck Jones, 1976; Fuhr, 1992) is perhaps the most well known classical probabilistic model. It assumes that terms are independently distributed in each of the two relevance models, so is essentially a naïve Bayes classifier for document ranking (Lewis, 1998). [1]

There have been several efforts to improve the binary representation. Van Rijsbergen extended the binary independence model by capturing some term dependency as defined by a minimum-spanning tree weighted by average mutual information (van Rijbergen, 1977). Croft (1981) investigated how the heuristic term significance weight can be incorporated into probabilistic models in a principled way. Another effort on improving document representation is to introduce the term frequency directly into the model by using a multiple 2-Poisson mixture representation of documents (Robertson et al., 1981). While this model has not shown superior empirical performance itself, an approximation of the model based on a simple TF formula turns out to be quite effective (Robertson and Walker, 1994). A different way of introducing term frequency into the model is implicit in text categorization approaches which view a document as being generated from a unigram language model (Kalt, 1996; McCallum and Nigam, 1998).

Models based on query generation ($p(D, Q \mid R) = p(Q \mid D, R)p(D \mid R)$) have been explored in (Maron and Kuhns, 1960), (Robertson et al., 1982), (Fuhr, 1992) and (Lafferty and Zhai, 2003). Indeed, the Probabilistic Indexing model proposed in (Maron and Kuhns, 1960) is the very first probabilistic retrieval model, in which the indexing terms assigned to a document are weighted by the probability that a user who likes the document would use the term in the query. That is, the weight of term t for document D is $p(t \mid D, r)$. However, the estimation of the model is based on user's feedback, not the content of D. The Binary Independence Indexing (BII) model proposed in (Fuhr, 1992) is another special case of the query generation model. It allows the description of a document (with weighted terms) to be estimated based on arbitrary queries, but the specific parameterization makes it difficult to estimate all the parameters in practice. In (Lafferty and Zhai, 2003), it has been shown that the recently proposed language modeling approach to retrieval can be viewed as a special probabilistic relevance model when query generation is used to decompose the generative model. This work provides a relevance-based justi-

[1] The required underlying independence assumption for the final retrieval formula is actually weaker (Cooper, 1991).

fication for this new family of probabilistic models based on statistical language modeling.

The language modeling approach was first introduced by Ponte and Croft (1998) and also explored in (Hiemstra and Kraaij, 1998; Miller et al., 1999; Berger and Lafferty, 1999; Song and Croft, 1999). The estimation of a language model based on a document (i.e., the estimation of $p(. \mid D, r)$) is the key component in the language modeling approach. Indeed, most work in this direction differs mainly in the language model used and the way of language model estimation. Smoothing of a document language model with some kind of collection language model has been very popular in the existing work. For example, geometric smoothing was used in (Ponte and Croft, 1998); linear interpolation smoothing was used in (Hiemstra and Kraaij, 1998; Berger and Lafferty, 1999), and was viewed as a 2-state hidden Markov model in (Miller et al., 1999). Berger and Lafferty explored "semantic smoothing" by estimating a "translation model" for mapping a document term to a query term, and reported significant improvements over the baseline language modeling approach through the use of translation models (Berger and Lafferty, 1999).

The language modeling approach has two important contributions. First, it introduces an effective probabilistic ranking function based on the query generation. While the earlier query generation models have all encountered difficulty in estimating the parameters, the model proposed in (Ponte and Croft, 1998) explicitly addresses the estimation problem through the use of statistical language models. Second, it reveals the connection between the difficult problem of text representation in IR and the language modeling techniques that have been well studied in other application areas such as statistical machine translation and speech recognition, making it possible to exploit various kinds of language modeling techniques to address the representation problem [2] .

Although the classic document generation probabilistic models and the language modeling approach can be seen as being based on the same notion of relevance and are probabilistically equivalent, they have several important differences from an estimation perspective, as they involve different parameters for estimation. When no relevance judgments are available, it is easier to estimate $p(Q \mid D, r)$ in the language modeling approach than to estimate $p(D \mid Q, r)$ in the classic probabilistic models. 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 BIR model has encountered difficulties in estimating $p(t \mid Q, r)$ and $p(t \mid Q, \overline{r})$ when no explicit relevance information is available. Typically, $p(t \mid Q, r)$ is set to a constant and $p(t \mid Q, \overline{r})$ is estimated under the assumption that the each document in the collection is not relevant (Croft and Harper, 1979; Robertson and Walker, 1997). Recently, Lavrenko and Croft made progress in estimating the rel-

[2] The use of a multinomial model for documents was actually first introduced in (Wong and Yao, 1989), but was not exploited as a language model.

evance model without relevance judgments by exploiting language modeling techniques (Lavrenko and Croft, 2001). On the other hand, when explicit relevance judgments are available, the classic models, being based on document generation, have the advantage of being able to naturally improve the estimation of the component probabilistic models by exploiting such explicit relevance information. This is because the relevance judgments from a user provide direct training data for estimating $p(t \mid Q, r)$ and $p(t \mid Q, \overline{r})$, which can then be applied to new documents. The same relevance judgments can also provide direct training data for improving the estimate of $p(t \mid D, r)$ in the language modeling approach, but only for those documents judged relevant. Thus, the directly improved models can not be expected to improve our ranking of other unjudged documents. Interestingly, such improved models can potentially be beneficial for new queries—a feature unavailable in document generation models.

Instead of imposing a strict document generation or query generation decomposition of the joint probability $p(D, Q \mid R)$, one can also "generate" a document-query pair simultaneously. Mittendorf and Schauble (1994) explored a passage-based generative model using Hidden Markov Model (HMM), which can be regarded as such a case. In this work, a document query pair is represented as a sequence of symbols, each corresponding to a term at a particular position of the document. All term tokens are clustered according to the similarity between the token and the query. In this way, a term token at a particular position of a document can be mapped to a symbol that represents the cluster the token belongs to. Such symbol sequences are modeled as the output from an HMM with two states, one corresponding to relevant passages and the other to the background noise. The relevance value is then computed based on the likelihood ratio of the sequence given the passage HMM model and the background model.

As seen below, probabilistic relevance models can be shown to be a special case of the risk minimization framework when a "constant cost" relevance-based loss function is used.

2.3 Probabilistic Inference Models

In a probabilistic inference model, the uncertainty of relevance of a document, with respect to a query, is modeled by the uncertainty associated with inferring the query from the document. Different inference models are possible depending on what it means to "infer a query from a document."

Van Rijsbergen introduced a logic-based probabilistic inference model for text retrieval (van Rijsbergen, 1986). In this model, a document is relevant to a query if (and only if) the query can be inferred from the document. The Boolean retrieval model can be regarded as a simple special case of this model. To cope with the

inherent uncertainty of relevance, van Rijsbergen introduced a logic for probabilistic inference, in which the probability of a conditional, such as $p \rightarrow q$, can be estimated based on the notion of possible worlds. Wong and Yao (1995) extended the probabilistic inference model and proposed a unified framework for supporting probabilistic inference with a concept space and a probability distribution defined over the concepts in the space. The probabilistic concept space model is shown to recover many other text retrieval models such as the Boolean, vector space, and the classic probabilistic models through different ways of modeling terms (thus document and query representations) in the concept space. Fuhr shows that some particular form of the language modeling approach can also be derived using this general probabilistic concept space model (Fuhr, 2001).

The inference network model is also based on probabilistic inference (Turtle and Croft, 1991). It is essentially a Bayesian belief network that models the dependency between the satisfaction of a query and the observation of documents. The estimation of relevance is based on the computation of the conditional probability that the query is satisfied given that the document is observed. Other similar uses of Bayesian belief network in retrieval have been presented in (Fung and Favero, 1995; Ribeiro and Muntz, 1996; Ribeiro-Neto et al., 2000). Kwok's network model may also be considered as performing a probabilistic inference (Kwok, 1995), though it is based on spread activation. The inference network model is a very general formalism; with different ways to realize the probabilistic relationship between the evidence of observing documents and the satisfaction of user's information need, one can obtain many different text retrieval models as special cases, including the Boolean, extended Boolean, vector space, and conventional probabilistic models. More importantly, it can potentially go beyond the traditional notion of topical relevance.

3 The Risk Minimization Framework

Informally, a retrieval system can be regarded as an interactive information service system that answers a user's query by presenting a list of documents. Usually the user would examine the presented documents and reformulate a query if necessary; the new query is then executed by the system to produce another new list of documents to present. At each iteration in this cycle, the retrieval system faces a decision-making problem – it needs to choose a subset of documents and present them to the user in some way, based on the available information to the system, which includes the current user, the user's query, the sources of documents, and a specific document collection. For example, the system may decide to select a subset of documents and present them without any particular order (as in Boolean retrieval); alternatively, it may decide to select all the documents and present them as a ranked list (as in the vector space model). In general, there could be many choices for the decision space, and we can regard the process of information re-

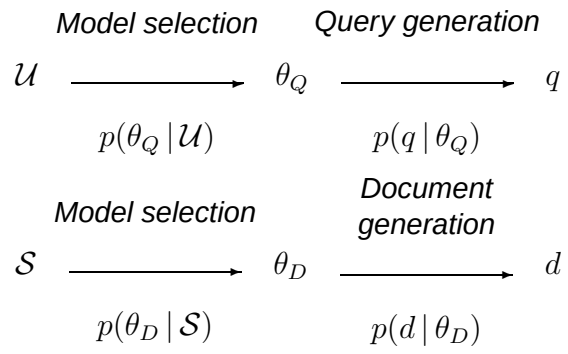

Fig. 1. Generative model of query q and document d.

trieval as consisting of a series of such decision making tasks.

We now formally define this decision problem. We view a query as being the output of some probabilistic process associated with the user $\mathcal{U}$, and similarly, we view a document as being the output of some probabilistic process associated with an author or document source $\mathcal{S}_i$. A query (document) is the result of choosing a model, and then generating the query (document) using that model. A set of documents is the result of generating each document independently, possibly from a different model. (The independence assumption is not essential, and is made here only to simplify the presentation.) 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 source or author.

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 q is then generated with probability $p(q \mid \theta_Q)$. Note that since a user can potentially use the same text query to mean different information needs, strictly speaking, the variable $\mathcal{U}$ should be regarded as corresponding to a user with the *current* context. Since this does not affect the presentation of the framework, we will simply refer to $\mathcal{U}$ as a user. Similarly, the source selects a document model θ_D according to a distribution $p(\theta_D \mid \mathcal{S})$, and then uses this model to generate a document d according to $p(d \mid \theta_D)$. Thus, we have Markov chains $\mathcal{U} \rightarrow \theta_Q \rightarrow q$ and $\mathcal{S} \rightarrow \theta_D \rightarrow d$. This is illustrated in Figure 1.

Let $\mathcal{C} = \{d_1, \ldots, d_N\}$ be a collection of documents obtained from sources $\vec{\mathcal{S}} = (\mathcal{S}_1, ..., \mathcal{S}_N)$. Our observations are thus $\mathcal{U}$, q, $\vec{\mathcal{S}}$, and $\mathcal{C}$. With this setup, we can now define retrieval actions. A retrieval action corresponds to a possible response of the system to a query. For example, one can imagine that the system would return an unordered subset of documents to the user. Alternatively, a system may decide a ranking of documents and present a ranked list of documents. Yet another possibility is to cluster the (relevant) documents and present a structured view of documents. Formally, a retrieval action can be defined as a compound decision

involving selecting a subset of documents D from $\mathcal{C}$ and presenting them to the user who has issued query q according to some presentation strategy π. Let Π be the set of all possible presentation strategies. We can represent all actions by $\mathcal{A} = \{(D, \pi)\}$, where $D \subseteq \mathcal{C}$ is a subset of $\mathcal{C}$ and $\pi \in \Pi$ is some presentation strategy.

In the general framework of Bayesian decision theory, to each such action $a = (D, \pi) \in \mathcal{A}$ there is associated a *loss* $L(a, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$, which in general depends upon all of the parameters of our model $\theta \equiv (\theta_Q, \{\theta_i\}_{i=1}^N)$ as well as any relevant user factors $F(\mathcal{U})$ and document source factors $F(\vec{\mathcal{S}})$, where θ_i is the model that generates document d_i. For convenience of notation, we will typically assume that the user factors $F(\mathcal{U})$ are included as part of the query model θ_Q, and similarly that the source factors $F(\vec{\mathcal{S}})$ are included as part of the document models θ_i; thus our loss function can be written as $L(a, \theta)$.

The *expected risk of action* a is given by

$$R(D, \pi \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C}) \;=\; \int_{\Theta} L(D, \pi, \theta)\, p(\theta \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C})\, d\theta$$

where the posterior distribution is given by

$$p(\theta \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C}) \;\propto\; p(\theta_Q \,|\, q, \mathcal{U}) \prod_{i=1}^{N} p(\theta_i \,|\, d_i, \vec{\mathcal{S}})$$

The Bayes decision rule is then to choose the action $\mathbf{a}^*$ having the least expected risk:

$$\mathbf{a}^* \;=\; (D^*, \pi^*) \;=\; \arg\min_{D,\pi} R(D, \pi \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C})$$

Thus, the document set D^* is selected and presented to the user with strategy π^*.

Note that this gives us a very general formulation of retrieval as a decision problem, which involves searching for D^* and π^* simultaneously. The presentation strategy can be fairly arbitrary in principle, e.g., presenting documents in a certain order, presenting a summary of the documents, or presenting a clustering view of the documents. However, we need to be able to quantify the loss associated with a presentation strategy.

We now consider several special cases of the risk minimization framework.

3.1 Set-based Retrieval

Let us consider the case when the loss function does not depend on the presentation strategy, which means that all we are concerned with is to select an optimal subset

of documents for presentation. In this case, the risk minimization framework leads to the following general set-based retrieval method.

$$\begin{aligned} D^* &= \arg\min_D R(D \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C}) \\ &= \arg\min_D \int_\Theta L(D, \theta)\, p(\theta \,|\, \mathcal{U}, q, \vec{\mathcal{S}}, \mathcal{C})\, d\theta \end{aligned}$$

The loss function can encode the user's preferences on the selected subset. Generally, the loss function will depend on the relevance status of the documents selected so that the optimal subset should contain the documents that are most likely to be relevant. But other preferences, such as the desired diversity and the desired size of a subset, can also be captured by an appropriate loss function.

The traditional Boolean retrieval model can be viewed as a special case of this general set-based retrieval framework, where we the uncertainty about the query models and document models is not modeled (e.g., $\theta_Q = q$ and $\theta_i = d_i$), and the following loss function is used:

$$L(D, \theta) = \sum_{d \in D} -\delta(d, q)$$

where $\delta(d, q) = 1$ if and only if document d satisfies the Boolean query q; otherwise $\delta(d, q) = -1$. This loss function is actually quite general, in the sense that if we allow $\delta(d, q)$ to be any deterministic retrieval rule applied to query q and document d, such that $\delta(d, q) > 0$ if d is relevant to q, otherwise $\delta(d, q) < 0$, then the loss function will always result in a retrieval strategy that involves making an independent binary retrieval decision for each document according to δ. In particular, the function δ can be defined on a structured query. One can easily imagine many other possibilities to specialize the set-based retrieval method.

3.2 Rank-based Retrieval

Let us now consider a different special case of the risk minimization framework where the selected documents are presented to the user as a ranked list of documents, so a possible presentation strategy corresponds to a possible ranking of documents. Such a ranking strategy has been assumed in most modern retrieval systems and models.

Formally, we may denote an action by $a = (D, \pi)$, where π is a complete ordering on D [3]. Taking action a would then mean presenting the selected documents in D one by one in the order given by π. This means that we can denote an action by a

[3] We could allow partial ordering in principle, but here we only consider complete ordering.

sequence of documents. So we will write $a = (d_{\pi(1)}, d_{\pi(2)}, ..., d_{\pi(k)})$, where $\pi(j)$ is the index of the document ranked at the j-th rank according to the permutation mapping π.

Let us further assume that our actions essentially involve different rankings of documents in the entire collection $\mathcal{C}$. That is, $\mathcal{A} = \{(\mathcal{C}, \pi)\}$, where π is a permutation over $[1..N]$, i.e., a complete ordering of the N documents in $\mathcal{C}$. To simplify our notation, we will use π to denote action $a = (\mathcal{C}, \pi)$.

In this case, the optimal Bayes decision is given by the following general ranking rule:

$$\begin{aligned}\pi^* &= \arg\min_{\pi} R(\pi \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\ &= \arg\min_{\pi} \int_{\Theta} L(\pi, \theta)\, p(\theta \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})\, d\theta\end{aligned}$$

where $\theta = (\theta_Q, \{\theta_i\}_{i=1}^N)$. We see that the loss function is now discriminating different possible rankings of documents.

How do we characterize the loss associated with a ranking of documents? Presenting documents by ranking implies that the user would apply some stopping criterion – the user would read the documents in order and stop wherever is appropriate. Thus, the actual loss (or equivalently utility) of a ranking would depend on where the user actually stops. That is, the utility is affected by the user's browsing behavior, which we could model through a probability distribution over all the ranks at which a user might stop. Given this setup, we can now define the loss for a ranking as the expected loss under the assumed "stopping distribution."

Formally, let s_i denote the probability that the user would stop reading after seeing the top i documents. We have $\sum_{i=1}^N s_i = 1$. We can treat $s_1, ..., s_N$ as user factors that depend on $\mathcal{U}$. Then the loss is given by

$$L(\pi, \theta) \;=\; \sum_{i=1}^{N} s_i\, \ell(\pi(1:i), \theta)$$

where $\ell(\pi(1:i), \theta)$ is the actual loss that would be incurred if the user actually views the first i documents according to π. Note that $L(\pi, \theta)$ and $\ell(\pi, \theta)$ are different: the former is the expected loss of the ranking under the user's "stopping probability distribution," while the latter is the exact loss of the ranking when the user actually views the whole list.

Assuming that the user would view the documents in the order presented, and the total loss of viewing i documents is the sum of the loss associated with viewing each individual document, we have the following reasonable decomposition of the

loss:

$$\ell(\pi(1:i),\theta) = \sum_{j=1}^{i} \ell(d_{\pi(j)} \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, \theta)$$

where $\ell(d_{\pi(j)} \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, \theta)$ is the conditional loss of viewing d_{π^j} given that the user has already viewed $(d_{\pi^1}, ..., d_{\pi^{j-1}})$.

Putting all of this together, we have

$$\begin{aligned}
\pi^* &= \arg\min_{\pi} R(\pi \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\
&= \arg\min_{\pi} \sum_{i=1}^{N} s_i \sum_{j=1}^{i} \int_{\Theta} \ell(d_{\pi(j)} \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, \theta)\, p(\theta \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})\, d\theta
\end{aligned}$$

Now, define the following conditional risk

$$r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \stackrel{\text{def}}{=} \int_{\Theta} \ell(d_k \mid d_1, ..., d_{k-1}, \theta)\, p(\theta \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})\, d\theta$$

which can be interpreted as the expected risk of the user's viewing document d_k given that $d_1, ..., d_{k-1}$ have been previously viewed. We can then write

$$\begin{aligned}
R(\pi \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) &= \sum_{i=1}^{N} s_i \sum_{j=1}^{i} r(d_{\pi(j)} \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\
&= \sum_{j=1}^{N} \left(\sum_{i=j}^{N} s_i \right) r(d_{\pi(j)} \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})
\end{aligned}$$

This is the general framework for ranking documents within the risk minimization framework. It basically says that the optimal ranking minimizes the expected conditional loss (under the stopping distribution) associated with sequentially viewing each document.

We see that the optimal ranking depends on the stopping distribution s_i. If a user tends to stop early, the optimal decision would be more affected by the loss associated with the top ranked documents; otherwise, it will be more equally affected by the loss associated with all the documents. Thus, the stopping probability distribution provides a way to model a "high-precision" (early stopping) preference or a "high-recall" (late stopping) preference. The sequential decomposition of the loss is reasonable when presenting a ranked list to the user. Clearly, when using other presentation strategies (e.g., clustering), such a decomposition would not be appropriate.

4 Loss Functions for Ranking

In this section we discuss specific loss functions, and show that the risk minimization framework includes several traditional retrieval models as special cases.

4.1 Independent Loss Functions

Let us first consider the case when the loss of viewing each document is independent of viewing others. That is,

$$\ell(d_{\pi(j)} \,|\, d_{\pi(1)}, ..., d_{\pi(j-1)}, \theta) \;=\; \ell(\theta_{\pi(j)}, \theta_Q)$$

which means

$$\ell(\pi(1:i), \theta) = \sum_{j=1}^{i} \ell(\theta_{\pi(j)}, \theta_Q)$$

In this case, the expected risk for ranking π is

$$\begin{aligned} R(\pi \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;&=\; \sum_{i=1}^{N} s_i \sum_{j=1}^{i} r(d_{\pi(j)} \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\ &= \sum_{j=1}^{N} \left(\sum_{i=j}^{N} s_i \right) r(d_{\pi(j)} \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \end{aligned}$$

We see that the risk of π is a weighted sum of the risk of viewing each individual document. As the rank increases, the weight decreases, with the weight on the first rank being the largest (i.e., $\sum_{i=1}^{N} s_i$). Thus, the optimal ranking π^*, independent of $\{s_i\}$, is in ascending order of the individual risk:

$$r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) = \int_{\Theta} \ell(d, \theta) p(\theta \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \, d\theta \tag{1}$$

This is equivalent to the situation where we assume a possible action is to present a *single* document. The loss function $\ell(d, \theta)$ can be interpreted as the loss associated with presenting/viewing document d, or equivalently the expected utility of presenting document d. Equation (1) thus specifies a general optimal ranking strategy which is very similar to the Probability Ranking Principle (Robertson, 1977); this connection will be further discussed in Section 6.

In general, there could be many different ways of specifying the loss function, which lead to different ranking functions. We now show that with appropriate choices of loss functions, many existing rank-based retrieval models can be derived in the risk minimization framework, including the vector space model, the classic probabilistic retrieval model, and the recently proposed language modeling

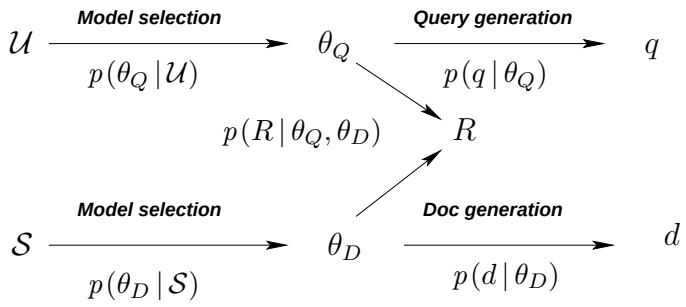

Fig. 2. Generative model of query q, document d, and relevance R.

approach. We also show that novel retrieval models, particularly those using statistical language models, can be systematically developed using the risk minimization framework.

4.1.1 Relevance-based loss functions

To show that the traditional relevance based probabilistic models are special cases of risk minimization, we consider the special case where the loss function L is defined through some binary relevance variable R. Specifically, we assume that for each document d_i, there is a hidden binary relevance variable R_i that depends on θ_Q and θ_i according to $p(R_i \mid \theta_Q, \theta_i)$, which is interpreted as representing the true relevance status of d_i with respect to q (1 for relevant and 0 for non-relevant); see Figure 2. The random variable R_i is observed when we have the user's relevance judgment on d_i, and is unobserved otherwise. Let us assume that R_i is not observed for now. Note that because the query model θ_Q can encode detailed knowledge about the user $\mathcal{U}$, the distribution of this relevance variable can be user-specific.

Introducing the variable R into our parameter space, equation (1) becomes:

$$r(d \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) = \sum_{R \in \{0,1\}} \int_{\Theta_D} \int_{\Theta_Q} \ell(R, \theta_D, \theta_Q)\, p(R \mid \theta_D, \theta_Q) p(\theta_D, \theta_Q \mid d, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})\, d\theta_Q\, d\theta_D \qquad (2)$$

Now let us assume that the loss function ℓ depends on θ only through the relevance variable R. That is, let ℓ be defined

$$\ell(R, \theta_D, \theta_Q) = \ell(R) = \begin{cases} c_0 \text{ if } R = 0 \\ c_1 \text{ if } R = 1 \end{cases}$$

where, c_0 and c_1 are two cost constants, and $c_0 > c_1$ for any reasonable loss function.

From equation (2), we have

$$
\begin{aligned}
r(d \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) &= c_0\, p(R = 0 \mid q, d) + c_1\, p(R = 1 \mid q, d) \\
&= c_0 + (c_1 - c_0)\, p(R = 1 \mid q, d)
\end{aligned}
$$

This means that the risk minimization ranking criterion is in this case equivalent to ranking based on $p(R = 1 \mid q, d)$, i.e., the probability of relevance given q and d[4]. This is the basis of all probabilistic relevance retrieval models. Thus, we have shown that the variants of the probabilistic relevance models reviewed in Section 2 are all special cases of the risk minimization framework. In particular, this includes both the classic document generation probabilistic retrieval models and the language modeling approach, which is based on query generation. (Lafferty and Zhai, 2001).

4.1.2 Proportional distance loss functions

Let us now consider a loss function ℓ which is proportional to a distance or similarity measure Δ between θ_Q and θ_D, i.e.,

$$
\ell(\theta_D, \theta_Q) \;=\; c\Delta(\theta_Q, \theta_D)
$$

where c is a constant cost. Intuitively, if the models θ, θ' are closer (more similar), then $\Delta(\theta, \theta')$ should be small, reflecting a user's goal of retrieving documents whose models are close to the query model.

With this loss function, from equation (1), we have

$$
r(d \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\propto\; \int_{\Theta_Q} \int_{\Theta_D} \Delta(\theta_Q, \theta_D)\, p(\theta_Q \mid q, \mathcal{U})\, p(\theta_D \mid d, \vec{\mathcal{S}}) d\theta_D\, d\theta_Q
$$

This means that the risk minimization ranking criterion is now equivalent to ranking based on the expected model distance. To make this distance easier to compute, we can approximate it by its value at the posterior mode of the parameters. That is,

$$
r(d \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\approx\; c\,\Delta(\widehat{\theta}_Q, \widehat{\theta}_D)
$$

where $\widehat{\theta}_Q = \arg\max_{\theta_Q} p(\theta_Q \mid q, \mathcal{U})$ and $\widehat{\theta}_D = \arg\max_{\theta_D} p(\theta_D \mid d, \vec{\mathcal{S}})$.

Note that the factor $p(\widehat{\theta}_D \mid d, \vec{\mathcal{S}})$ 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. Thus we see that under these assumptions and approximations, $r(d \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\propto\; \Delta(\widehat{\theta}_D, \widehat{\theta}_Q)$ We can view the vector space

[4] Since $c_0 > c_1$, a decreasing order in r is equivalent to an increasing order in $p(R = 1 \mid q, d)$.

model as a special case of this general similarity model, in which $\widehat{\theta}_Q$ and $\widehat{\theta}_D$ are simply term vector parameters estimated heuristically and the distance function is the cosine or inner product measure.

As a special case of the distance-based model, we assume that θ_Q and θ_D are the parameters of unigram language models, and choose as the distance function the Kullback-Leibler divergence. This leads to

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

and

$$\begin{aligned} r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;&\propto\; -\sum_w p(w \,|\, \widehat{\theta}_q) \log p(w \,|\, \widehat{\theta}_d) + \textit{cons}(q) \\ &\stackrel{\text{rank}}{=} -\sum_w p(w \,|\, \widehat{\theta}_q) \log p(w \,|\, \widehat{\theta}_d) \end{aligned}$$

Thus the ranking function is essentially the cross entropy of the query language model with respect to the document language model. The dropped constant is minus the query model entropy. The value of the cross entropy is always larger than or equal to the query model entropy. The minimum value (i.e. query model entropy) is achieved when $\widehat{\theta}_d$ is identical to $\widehat{\theta}_q$, which makes sense for retrieval.

The KL-divergence model covers the popular query likelihood ranking function as a special case. Indeed, suppose $\widehat{\theta}_q$ is just the empirical distribution of the query $q = (q_1, q_2, \ldots, q_m)$, That is, $\widehat{\theta}_q$ is the language model given by

$$p(w \,|\, \widehat{\theta}_q) \;\;=\;\; \frac{1}{m} \sum_{i=1}^{m} \delta(w, q_i)$$

where $\delta(w, q_i)$ is the indicator function. We obtain

$$r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\;\propto\;\; -\frac{1}{m} \sum_{i=1}^{m} \log p(q_i \,|\, \widehat{\theta}_d)$$

This is precisely the log-likelihood criterion used by Ponte and Croft (1998) in introducing the language modeling approach, which has been used in all work on the language modeling approach to date. In (Zhai and Lafferty, 2001), new methods were developed to estimate a model $\widehat{\theta}_Q$, leading to significantly improved performance over the use of the empirical distribution $\widehat{\theta}_Q$.

4.1.3 "Binned" distance loss functions

We now consider another special loss function based on a distance function, indexed by a small constant ϵ:

$$\ell_\epsilon(\theta_D, \theta_Q) \;=\; \begin{cases} 0 \text{ if } \Delta(\theta_Q, \theta_D) \leq \epsilon \\ c \text{ otherwise} \end{cases}$$

where $\Delta : \Theta_Q \times \Theta_D \to \mathbb{R}$ is a model distance function, and c is a constant positive cost. Thus, the loss is zero when the query model and the document model are close to each other, and is c otherwise, capturing a user's preference for retrieving documents whose models are close to the query model.

We can show that this loss function leads to a family of two-stage language models explored in (Zhai and Lafferty, 2002). First, we see that the risk is

$$r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) = c - \int_{\Theta_D} \int_{\theta_Q \in S_\epsilon(\theta_D)} p(\theta_Q \,|\, q, \mathcal{U})\, p(\theta_D \,|\, d, \vec{\mathcal{S}})\, d\theta_Q\, d\theta_D$$

where $S_\epsilon(\theta_D) = \{\theta_Q \,|\, \Delta(\theta_Q, \theta_D) < \epsilon\}$.

Now, assuming that $p(\theta_D \,|\, d, \vec{\mathcal{S}})$ is concentrated on an estimated value $\hat{\theta}_D$, we can approximate the value of the integral over Θ_D by the integrand's value at $\hat{\theta}_D$. Note that the constant c can be ignored for the purpose of ranking. Thus, using the notation $A \overset{\text{rank}}{\approx} B$ to mean that A and B have the same effect for ranking, we have that

$$r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\overset{\text{rank}}{\approx}\; -\int_{\theta_Q \in S_\epsilon(\hat{\theta}_D)} p(\theta_Q \,|\, q, \mathcal{U})\, d\theta_Q$$

When θ_Q and θ_D belong to the same parameter space (i.e., $\Theta_Q = \Theta_D$) and ϵ is very small, the value of the integral can be approximated by the value of the function at $\hat{\theta}_D$ times a constant (the volume of $S_\epsilon(\hat{\theta}_D)$), and the constant can again be ignored for the purpose of ranking. That is,

$$r(d \,|\, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \;\overset{\text{rank}}{\approx}\; -p(\hat{\theta}_D \,|\, q, \mathcal{U})$$

Therefore, using this loss we will be actually ranking documents according to $p(\hat{\theta}_D \,|\, q, \mathcal{U})$, i.e., the posterior probability that the user used the estimated document model as the query model. Applying Bayes' formula, we can rewrite this as

$$p(\hat{\theta}_D \,|\, q, \mathcal{U}) \;\propto\; p(q \,|\, \hat{\theta}_D, \mathcal{U}) p(\hat{\theta}_D \,|\, \mathcal{U}) \tag{3}$$

Equation (3) is the basic two-stage language model retrieval formula, in which $p(q \mid \hat{\theta}_D, \mathcal{U})$ captures how well the estimated document model $\hat{\theta}_D$ explains the query, whereas $p(\hat{\theta}_D \mid \mathcal{U})$ encodes our prior belief that the user would use $\hat{\theta}_D$ as the query model. It can also be regarded as a natural generalization of the basic language modeling approach (i.e., the simple query likelihood method). In (Zhai and Lafferty, 2002) this two-stage language model is shown to achieve excellent retrieval performance through completely automatic setting of parameters.

4.2 Dependent Loss Functions

We have demonstrated how the risk minimization framework can recover existing retrieval models and can motivate some interesting new retrieval models through independent loss functions. However, an independent loss function is rarely an accurate model of real retrieval preferences; the loss of viewing one document generally depends on the documents already viewed. For example, if the user has already seen the same document or a similar document, then the document should incur a much greater loss than if it were completely new to the user. In this section, we discuss dependent loss functions.

When an independent loss function is used, we can derive the exact optimal ranking strategy (i.e., equation (1)) which does not depend on the stopping probability distribution and can be computed efficiently. However, when a dependent loss function is used , the complexity of finding the optimal ranking makes the computation intractable. One practical solution is to use a greedy algorithm to construct a sub-optimal ranking. Specifically, we can "grow" the target ranking by choosing the document at each rank, starting from the very first rank. Suppose we already have a partially constructed ranking $\pi(1:i)$, and we are now choosing the document at rank $i+1$. Let k be a possible document index to be considered for rank $i+1$, and let $\pi(1:i,k)$ represent the ordering $(d_{\pi(1)}, ..., d_{\pi(i)}, d_k)$. Then, the increase in risk due to choosing d_k at rank $i+1$ is

$$\begin{aligned}
\delta(k \mid \pi(1:i)) &= R(\pi(1:i,k) \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) - R(\pi(1:i) \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\
&= s_{i+1}\,(r(d_k \mid d_{\pi(1)}, ..., d_{\pi(i)}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) + \\
&\quad \sum_{j=1}^{i} r(d_j \mid d_{\pi(1)}, ..., d_{\pi(j-1)}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}))
\end{aligned}$$

Thus, at each step we just need to evaluate

$$\delta^{'}(k \mid \pi(1:i)) = r(d_k \mid d_{\pi(1)}, ..., d_{\pi(i)}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$$

and choose the k that minimizes $\delta^{'}(k \mid \pi(1:i))$.

This gives us a general greedy and context-dependent ranking algorithm. Interestingly, due to the use of a greedy strategy, we see again that the "optimal" ranking does not depend on the stopping probabilities s_i. In the next section, we discuss how we may instantiate this general algorithm with specific dependent loss functions in the context of a non-traditional ranking task—subtopic retrieval.

5 Models for Subtopic Retrieval

5.1 The problem of subtopic retrieval

A regular retrieval task is often framed as the problem of retrieving relevant documents based on the assumption that a single document is the information unit under consideration. However, a topic usually has some degree of subtopic structure. For example, a student doing a literature survey on "machine learning" may be most interested in finding documents that cover representative approaches to the field and the relations between these approaches. If a topic often has a unique structure that involves many different subtopics, a user with a high recall retrieval preference may prefer a ranking of documents where the top documents cover different subtopics. This problem, referred to as "aspect retrieval," was investigated in the TREC interactive track (Over, 1998), where the purpose was to study how an interactive retrieval system can help a user to efficiently gather diverse information about a topic.

How can we formally define a retrieval model for such a subtopic retrieval problem? Clearly, this requires non-traditional ranking of documents, since ranking solely based on relevance would not be optimal. We thus need non-traditional ranking models that can not only model relevance but also model redundancy, novelty, and subtopics. To model the subtopic retrieval task in the risk minimization framework we require a dependent loss function. In this section we present two different types of dependent loss functions that are appropriate for this task.

The first type of loss function is the Maximal Marginal Relevance (MMR) loss function, in which we encode a preference for retrieving documents that are both topically relevant and novel (Carbonell and Goldstein, 1998). In essence, the goal is to retrieve relevant documents and, at the same time, minimize the chance that the user will see redundant documents as he or she goes through the ranked list of documents. Intuitively, as we reduce the redundancy among documents, we can expect the coverage of the same subtopic to be minimized and thus the coverage of potentially different subtopics to increase.

The second type of loss function is the Maximal Diverse Relevance (MDR) loss function, in which we encode a preference for retrieving documents that best sup-

plement the previously retrieved documents, in terms of covering different subtopics. We thus need to model both topical relevance and subtopic structure of documents. Intuitively, an MDR loss function will assess which subtopics have been well covered and which are under covered, and then prefer a document that best treats those under-covered subtopics. We now discuss both types of dependent loss functions in detail.

5.2 Maximal Marginal Relevance (MMR) Loss Functions

The idea of Maximal Marginal Relevance (MMR) ranking was first proposed and formalized in (Carbonell and Goldstein, 1998). It is based on the assumption that one should consider not only the relevance value, but also the novelty (or equivalently, redundancy) in the presented documents. Informally, given a set of previously selected documents, the next best document is one that is both relevant to the query topic and different from the already selected documents. In the risk minimization framework, we can encode such preferences with a conditional loss function $\ell(d_k \mid d_1, ..., d_{k-1}, \theta)$ that "balances" the relevance value and the redundancy value of a document.

If $\ell_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \theta_1, ..., \theta^k)$ is such a loss function, the conditional risk is then

$$
\begin{aligned}
& r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \; = \\
& = \int_{\Theta} l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \theta_1, ..., \theta_k) p(\theta \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \, d\theta
\end{aligned}
$$

If we assume that the parameters θ are concentrated at the mode $\hat{\theta} \equiv (\hat{\theta_Q}, \{\hat{\theta}_i\}_{i=1}^k)$, then the posterior distribution is close to a delta function. In this simplified case, ranking based on the conditional risk is approximately equivalent to ranking based on the value of the loss function at the mode, i.e.,

$$
r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \quad \overset{\text{rank}}{\approx} \quad l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \hat{\theta}_Q, \hat{\theta}_1, ..., \hat{\theta}_k)
$$

An MMR loss function requires the combination of a relevance measure and a novelty measure. While there may be many different ways to specify such a loss function, the problem of deriving a well motivated loss of this type largely remains an open research question (Zhai, 2002).

Suppose we make the simplifying assumption that a relevance score and a novelty score can be computed independently. In this case we can define our loss function as a direct combination of the two scores. Let $S_R(\theta_k; \theta_Q)$ be any relevance scoring function and $S_N(\theta_k; \theta_1, ..., \theta_{k-1})$ any novelty scoring function. An MMR loss

function can then be defined as a combination of the two scoring functions as

$$l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \{\theta_i\}_1^{k-1}) = f(S_R(\theta_k; \theta_Q), S_N(\theta_k; \theta_1, ..., \theta_{k-1}), \mu)$$

where $\mu \in [0, 1]$ is a relevance-novelty trade-off parameter, such that

$$l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \theta_1, ..., \theta_{k-1}) \stackrel{\text{rank}}{=} \begin{cases} S_R(\theta_k; \theta_Q) & \text{if } \mu = 0 \\ S_N(\theta_k; \theta_1, ..., \theta_{k-1}) & \text{if } \mu = 1 \end{cases}$$

One such combination is the linear interpolation of S_R and S_N, given by

$$l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \{\theta_i\}_1^{k-1}) = (1 - \mu) S_R(\theta_k; \theta_Q) + \mu S_N(\theta_k; \theta_1, ..., \theta_{k-1})$$

which is precisely the original MMR formula presented in (Carbonell and Goldstein, 1998). Clearly, this loss function makes sense only when the range of the functions S_R and S_N are comparable.

When relevance and novelty/redundancy are computed with a probabilistic model, we can use the following general loss function:

$$\begin{aligned} l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \{\theta_i\}_1^{k-1}) &= c_1 p(\text{Rel} \mid d) p(\text{New} \mid d) \\ &+ c_2 p(\text{Rel} \mid d)(1 - p(\text{New} \mid d)) \\ &+ c_3 (1 - p(\text{Rel} \mid d) p(\text{New} \mid d) \\ &+ c_4 (1 - p(\text{Rel} \mid d))(1 - p(\text{New} \mid d)) \end{aligned}$$

where c_1, c_2, c_3, and c_4 are cost constants; $p(\text{Rel} \mid d)$ is the probability that document d is relevant; and $p(\text{New} \mid d)$ is the probability that d is new with respect to documents $d_1, ..., d_{k-1}$.

We may reasonably assume that $c_3 = c_4$, since whether or not a non-relevant document carries new information is presumably not interesting to the user. We can also reasonably assume that there is no cost incurred if the document is both relevant and (completely) new, i.e., $c_1 = 0$. Under these two assumptions, we have

$$\begin{aligned} &l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \{\theta_i\}_1^{k-1}) = \\ &\quad c_2 p(\text{Rel} \mid d)(1 - p(\text{New} \mid d)) + c_3 (1 - p(\text{Rel} \mid d) \end{aligned}$$

For any reasonable loss function, both c_2 and c_3 should be some positive cost, and usually $c_3 > c_2$. In general, c_2 and c_3 may change according to k, or even the actual documents $d_1, ..., d_{k-1}$. Intuitively, c_2 is the cost of seeing a relevant, but redundant document, whereas c_3 is the cost of seeing a non-relevant document. Clearly, when $c_2 = 0$, so that the user is assumed not to care about redundancy, the loss function is

based on the probability of relevance. We assume below that $c_2 > 0$, which allows us to re-write the loss function in the following equivalent form for the purpose of ranking documents:

$$\begin{aligned} l_{MMR}(d_k \mid d_1, ..., d_{k-1}, \theta_Q, \{\theta_i\}_1^{k-1}) &= c_3 + c_2 p(\text{Rel} \mid d)(1 - \frac{c_3}{c_2} - p(\text{New} \mid d)) \\ &\stackrel{\text{rank}}{=} p(\text{Rel} \mid d)(1 - \frac{c_3}{c_2} - p(\text{New} \mid d)) \end{aligned}$$

Note that a higher $p(\text{New} \mid d)$ always helps to reduce the loss, and when $c_2/c_2 \geq 1$, a higher $p(\text{Rel} \mid d)$ also implies a smaller loss. However, reduction in loss affected by the cost ratio c_3/c_2, which indicates the relative cost of seeing a non-relevant document compared with seeing a relevant but redundant document. When the ratio is large, i.e., $c_3 \gg c_2$, the influence of $p(\text{New} \mid d)$ could be negligible. This means that when the user has low tolerance for any non-relevant document, the optimal ranking would essentially be relevance-based, and not affected by the novelty of documents. When $c_3 = c_2$, we would score documents based on $p(\text{Rel} \mid d)p(\text{New} \mid d)$, which is essentially the scoring formula for generating temporal summaries proposed in (Allan et al., 2001), where $p(\text{Rel} \mid d)$ is denoted $p(\text{Useful} \mid d)$. In practice, there will be a compromise between retrieving documents with new content and avoiding non-relevant documents. In (Zhai, 2002; Zhai et al., 2003), this loss function is investigated with $p(\text{Rel} \mid d)$ being assumed to be proportional to $p(q \mid d)$ and $p(\text{New} \mid d)$ being estimated with a mixture language model.

A deficiency in way the MMR loss function combines the relevance score and the novelty score lies in the assumption of independent relevance and novelty. In other words, one does not have a direct measure of relevance of the novel information contained in a new document. Thus, a document formed by concatenating a previously seen (and therefore redundant) relevant document with new, but irrelevant information may be ranked highly, even though it is useless to the user. Several alternative MMR loss functions that directly measure the relevance of the new information are explored in (Zhai, 2002).

5.3 Maximal Diverse Relevance (MDR) Loss Functions

We now discuss a different type of loss function for the subtopic retrieval task. MMR loss functions aim to increase the subtopic coverage indirectly through eliminating the redundancy among documents. Here we the goal is to improve the subtopic coverage more directly by modeling the possible subtopics in the documents.

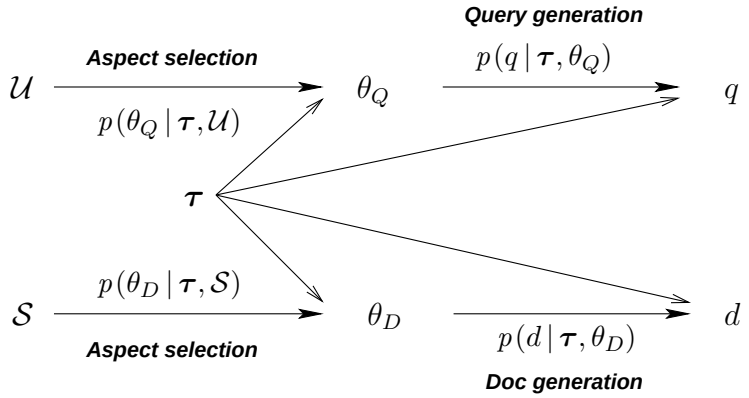

Fig. 3. Aspect generative model of query q and document d.

5.3.1 A General Subtopic Retrieval Model

To model the subtopics, we consider the generative model illustrated in Figure 3. We assume that there is a space of A subtopics, each characterized by a unigram language model. Formally, let $\boldsymbol{\tau} = (\tau_1, ..., \tau_A)$ be a vector of subtopics, where τ_i is a unigram language model and $p(w \mid \tau_i)$ gives the probability of word w according to the subtopic τ_i.

Now, let us assume that a user, with an interest in retrieving documents to cover some of these A subtopics, would first pick a probability distribution θ_Q over the subtopics, and then formulate a query according to a query generation model $p(q \mid \boldsymbol{\tau}, \theta_Q)$. Intuitively, θ_Q encodes preferences on subtopic coverage, and in general, would have the probability mass concentrated on those subtopics that are most interesting to the user.Furthermore, among these "interesting subtopics," the distribution is generally non-uniform, reflecting the fact that some subtopics are more important than others. Similarly, we also assume that the author or source of a document d would first pick a subtopic coverage distribution θ_D, and then generate d according to a document generation model $p(d \mid \boldsymbol{\tau}, \theta_D)$. A simple example of such a model $p(d \mid \boldsymbol{\tau}, \theta_D)$ would be a mixture model, in which θ_D is the mixing weights and $\boldsymbol{\tau}$ are the component unigram language models. That is, with $d = d_1 d_2 ... d_n$, we have

$$p(d \mid \boldsymbol{\tau}, \theta_D) = \prod_{i=1}^{n} \sum_{j=1}^{A} p(j \mid \theta_D) p(d_i \mid \tau_j)$$

However, the derivation below is not restricted to such a mixture model.

To derive a subtopic retrieval model, we start with the following general greedy ranking formula:

$$r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \stackrel{\text{def}}{=} \int_{\Theta} \ell(d_k \mid d_1, ..., d_{k-1}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))\, p(\theta \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})\, d\theta$$

This conditional risk gives us a way to evaluate the remaining documents and pick the best d_k, given that we have already selected d_1, ..., d_{k-1}. With the generative models given above, $\theta = (\boldsymbol{\tau}, \theta_Q, \theta_{D_1}, ..., \theta_{D_k})$.

We now consider the following loss function:

$$\begin{aligned}\ell(d_k \,|\, d_1, ..., d_{k-1}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) &= \ell(d_k \,|\, d_1, ..., d_{k-1}, \boldsymbol{\tau}, \theta_Q, \theta_{D_1}, ..., \theta_{D_k}) \\ &= D(\theta_Q || \, \theta^{D_k}_{D_1 ... D_{k-1}})\end{aligned}$$

where $\theta^{D_k}_{D_1 ... D_{k-1}}$ is a weighted average of $\{\theta_{d_i}\}_{i=1}^{k}$ defined as follows:

$$p(a \,|\, \theta^{D_k}_{D_1 ... D_{k-1}}) = \frac{\mu}{k-1} \sum_{i=1}^{k-1} p(a \,|\, \theta_{D_i}) + (1-\mu) p(a \,|\, \theta_{D_k})$$

where $\mu \in (0, 1]$ is a parameter indicating how much redundancy we would like to model.

The idea behind this loss function is that we expect θ_Q to indicate which subtopics are relevant—a high $p(a \,|\, \theta_Q)$ indicates that the subtopic a is likely a relevant one. The loss function encodes our preferences for a similar "subtopic coverage distribution" given by all the documents $d_1, ..., d_k$. Thus, if θ_Q assigns high probabilities to some subtopics, then we would expect to cover these (presumably relevant) subtopics more than other subtopics. The best d_k is thus the one that can work together with $d_1, ..., d_{k-1}$ to achieve a coverage distribution that is most similar to the desired subtopic coverage based on the query, i.e., $p(a \,|\, \theta_Q)$. The parameter μ controls how much we rely on the previously chosen documents $d_1, ..., d_{k-1}$ to cover the subtopics. If we do not rely on them (i.e., $\mu = 0$), we will be looking for a d_k that best covers all the relevant subtopics by itself. On the other hand, if $\mu > 0$, part of the coverage would have been explained by the previously chosen documents, and the best d_k would be one that best covers those "under-covered" relevant subtopics. Essentially, we are searching for the d_k that best supplements the coverage provided by the previously selected documents with respect to the desired coverage θ_Q.

Putting this loss function and the subtopic generative model into the conditional risk formula, we have

$$\begin{aligned}&r(d_k \,|\, d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\ &= \int_\Theta \ell(d_k \,|\, d_1, ..., d_{k-1}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) p(\theta \,|\, q, \mathcal{C}, \mathcal{C}_k, \mathcal{U}, \vec{\mathcal{S}}) \, d\theta \\ &= \int_\Theta D(\theta_Q || \, \theta^{D_k}_{D_1 ... D_{k-1}}) p(\theta_Q, \theta_{D_1}, ..., \theta_{D_k} \,|\, q, \mathcal{C}, \mathcal{C}_k, \mathcal{U}, \vec{\mathcal{S}}) d\theta_Q d\theta_{D_1} ... d\theta_{D_k}\end{aligned}$$

and

$$
\begin{aligned}
p(\theta_Q, \theta_{D_1}, ..., \theta_{D_k} \mid q, \mathcal{C}, \mathcal{C}_k, \mathcal{U}, \vec{\mathcal{S}}) &= \int_{\mathcal{T}} p(\theta_Q, \theta_{D_1}, ..., \theta_{D_k}, \boldsymbol{\tau} \mid q, \mathcal{C}, \mathcal{C}_k, \mathcal{U}, \vec{\mathcal{S}}) d\boldsymbol{\tau} \\
&= \int_{\mathcal{T}} p(\theta_Q \mid \boldsymbol{\tau}, q, \mathcal{U}) \prod_{i=1}^{k} p(\theta_{D_i} \mid \boldsymbol{\tau}, d_i, \vec{\mathcal{S}}) p(\boldsymbol{\tau} \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) d\boldsymbol{\tau} \\
&\approx p(\theta_Q \mid \hat{\boldsymbol{\tau}}, q, \mathcal{U}) \prod_{i=1}^{k} p(\theta_{D_i} \mid \hat{\boldsymbol{\tau}}, d_i, \vec{\mathcal{S}}) p(\hat{\boldsymbol{\tau}} \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})
\end{aligned}
$$

where $\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} p(\boldsymbol{\tau} \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$, and $\mathcal{C}_k = \{d_1, ..., d_k\}$.

Note that we have assumed that $\boldsymbol{\tau}$ can be estimated using all the documents in the collection, so $p(\boldsymbol{\tau} \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$ does not depend on d_k and can be ignored for the purpose of ranking d_k. That is,

$$
\begin{aligned}
& r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) \\
& \quad \stackrel{\text{rank}}{\approx} \int_{\Theta} D(\theta_Q || \, \theta^{D_k}_{D_1 ... D_{k-1}}) p(\theta_Q \mid \hat{\boldsymbol{\tau}}, q, \mathcal{U}) \prod_{i=1}^{k} p(\theta_{D_i} \mid \hat{\boldsymbol{\tau}}, d_i, \vec{\mathcal{S}}) \\
& \quad \stackrel{\text{rank}}{\approx} D(\widehat{\theta}_Q || \, \widehat{\theta}^{D_k}_{D_1 ... D_{k-1}}) p(\widehat{\theta}_Q \mid \hat{\boldsymbol{\tau}}, q, \mathcal{U}) \prod_{i=1}^{k} p(\widehat{\theta}_{D_i} \mid \hat{\boldsymbol{\tau}}, d_i, \vec{\mathcal{S}}) \\
& \quad \stackrel{\text{rank}}{=} D(\widehat{\theta}_Q || \, \widehat{\theta}^{D_k}_{D_1 ... D_{k-1}}) p(\widehat{\theta}_{D_k} \mid \hat{\boldsymbol{\tau}}, d_k, \vec{\mathcal{S}}) \\
& \quad \stackrel{\text{rank}}{\approx} D(\widehat{\theta}_Q || \, \widehat{\theta}^{D_k}_{D_1 ... D_{k-1}})
\end{aligned}
$$

where $\widehat{\theta}_Q = \arg\max_{\theta_Q} p(\theta_Q \mid \hat{\boldsymbol{\tau}}, q, \mathcal{U})$ and $\widehat{\theta}_{D_i} = \arg\max_{\theta_{D_i}} p(\theta_{D_i} \mid \hat{\boldsymbol{\tau}}, d_i, \vec{\mathcal{S}})$. Thus, we have obtained the following ranking procedure:

(1) Estimate $\boldsymbol{\tau}$, i.e., $\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} p(\boldsymbol{\tau} \mid q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$, before selecting any document.
(2) Rank all the documents in a greedy fashion, using the conditional risk $r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$ when selecting the k-th document.
(3) Compute $r(d_k \mid d_1, ..., d_{k-1}, q, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}})$ by first computing $\widehat{\theta}_Q$ and $\widehat{\theta}_{D_k}$ and then evaluate $D(\widehat{\theta}_Q || \, \widehat{\theta}^{D_k}_{D_1 ... D_{k-1}})$.

In order to make this general subtopic retrieval model operational, we need to specify a query model ($p(q \mid \boldsymbol{\tau}, \theta_Q)$ and $p(\theta_Q \mid \boldsymbol{\tau}, \mathcal{U})$) and a document model ($p(d \mid \boldsymbol{\tau}, \theta_D)$ and $p(\theta_D \mid \boldsymbol{\tau}, \vec{\mathcal{S}})$).

In general, we can plug in any specific subtopic-based generative models to the general subtopic retrieval model, leading to potentially different retrieval formulas. For example, the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) has been explored in (Zhai, 2002).

6 Discussion

6.1 A Decision-Theoretic View of Retrieval

Treating retrieval from a decision-theoretic view is not new; in the 1970s, researchers were already studying how to choose and weight indexing terms from a decision-theoretic perspective (Bookstein and Swanson, 1975; Harter, 1975; Cooper and Maron, 1978). The probability ranking principle had also been justified based on optimizing the statistical decision about whether to retrieve a document (Robertson, 1977). However, the action/decision space considered in all this early work was limited to a binary decision regarding whether to retrieve a document or assign an index term to a document.

In the risk minimization framework, we have explicitly and formally treated the retrieval problem as a decision-making problem. The decision problem is a more general one where the action space, in principle, consists of all the possible actions that the system can take in response to a query. The scope of the decision space is a significant departure from existing decision-theoretic treatment of retrieval (Wong et al., 1991; Dominich, 2001). Such a general decision-theoretic view explicitly suggests that retrieval is modeled as an *interactive* process that involves cycles of a user reformulating the query and the system presenting information. Indeed, a user variable ($\mathcal{U}$) and a document source variable ($\mathcal{S}$) have been explicitly and formally introduced into the retrieval models in order to allow this level of generality.

A difference between the risk minimization framework and the early decision-theoretic treatment of indexing is that the early work, such as (Cooper and Maron, 1978), uses utility in a frequency sense, i.e., the expected utility over *all possible future uses*, whereas we take a Bayesian view and consider the utility with respect to the *current* user and the available evidence. The decision-theoretic view of retrieval allows the risk minimization framework to be more general than other retrieval frameworks such as the probabilistic inference framework proposed in (Wong and Yao, 1995) and the inference network framework (Turtle and Croft, 1991).

6.2 Risk Minimization and the Probability Ranking Principle

The Probability Ranking Principle (PRP) has often been taken as the foundation for probabilistic retrieval models. As stated in (Robertson, 1977), the principle is based on the following two assumptions:

(a) The *relevance* of a document to a request is independent of the other documents in the collection;
(b) The *usefulness* of a relevant document to a requester may depend on the

number of relevant documents the requester has already seen (the more he has seen, the less useful a subsequent one may be).

Under these assumptions, the PRP provides a justification for ranking documents in descending order of probability of relevance, which can be evaluated separately for each document.

Using the risk minimization framework, we have derived a general ranking formula for ranking documents based on an ascending order of the expected risk of a document, which can also be computed separately for each document. And we have also made two assumptions:

(a) *Independent loss*. The loss associated with a user's viewing of one document does not depend on any other documents that the user may have seen.
(b) *Sequential browsing*. When presented with a ranked list of documents, a user will browse through the list sequentially according to the ranking.

It is interesting to note the relationship between these two assumptions and the two assumptions made in (Robertson, 1977). The sequential browsing assumption is also made in (Robertson, 1977), though it is not explicitly stated, but our independent loss assumption is stronger than the independent relevance assumption, since it is possible to define a dependent loss function based on independent relevance. Indeed, the second assumption in (Robertson, 1977) implies that the utility (or equivalently, the loss) of retrieving one document depends on the number of relevant documents that are ranked above this document, though it does not directly depend on the relevance status of any specific document. The price for this weaker assumption, however, is that the PRP is no longer guaranteed to give a ranking that is optimal globally, but only one that is optimal as a greedy algorithm. The assumption that a greedy algorithm is used to construct the optimal ranking is implicit in (Robertson, 1977), since the decision problem involves retrieving a single document rather than choosing a ranking of all documents. In contrast, under our assumptions, ranking based on the expected risk can be shown to be globally optimal.

The PRP has several limitations as discussed in, e.g., (Cooper, 1994). First, it assumes that document usefulness is a binary property, but in reality it should really be a matter of degree. The independent loss ranking function that we derived does not have this limitation. Indeed, it is possible to derive the PRP in the risk minimization framework by assuming that the loss function depends only on a binary relevance variable. Second, a ranking of documents by probability of usefulness is not always optimal. Cooper gave such an example, which essentially shows that the independent relevance assumption may not be true. Robertson discussed informally two ways to extend the PRP to address the possible dependency among documents (Robertson, 1977). Both have been captured in the risk minimization framework. The first is to go from ranking based on probability of relevance to ranking based on

expected utility, which we achieve by using a loss function in the risk minimization framework. The second is essentially the greedy algorithm for ranking based on the conditional loss function. Thus, in the risk minimization framework we provide a formal way to go beyond the PRP. As stated in (Robertson, 1977),

> The estimation of probability of relevance for each document may not be the most appropriate form of prediction. The two main questions are:
> - On the basis of what kinds of information can the system make the prediction?
> - How should the system utilize and combine these various kinds of information?
>
> These questions represent, indeed, the central problem of retrieval theory.

The risk minimization framework provides a formal answer to both of the questions. The information available to the system includes the user ($\mathcal{U}$), the document source ($\vec{\mathcal{S}}$), the query (q), and the documents ($\mathcal{C}$). A "prediction" consists of selecting a subset of documents and presenting them in some way. However, one can easily imagine other possible "predictions." These factors are combined in a Bayesian decision theoretic framework to compute an optimal prediction.

6.3 The Notion of Relevance

The risk minimization framework was originally motivated by the need for a general ranking procedure that allows one to view several different ranking criteria, including the query-likelihood criterion used in the language modeling approach, within the same unified framework. As discussed in the existing literature, the retrieval problem may be decomposed into three basic components: representation of a query, representation of a document, and matching the two representations. With an emphasis on the implementation of the framework and probabilistic modeling, we make three corresponding assumptions: (1) A query can be viewed as an observation from a probabilistic query model; (2) A document can be viewed as an observation from a probabilistic document model; (3) The utility of a document with respect to a query (i.e., the ranking criterion) is a function of the query model and document model. Flexibility in choosing different query models and document models is necessary to allow different representations of queries and documents. The flexibility of choosing the loss function is necessary in order to cover different notions of relevance and different ranking strategies.

As a result of these assumptions, the representation problem is essentially equivalent to that of model estimation, while the matching problem is equivalent to the estimation of the value of a utility function based on the observed query and document. In Bayesian decision theory, utility is modeled by a loss function; a loss value can be regarded as a negative utility value. Thus, we can say that the notion of relevance taken in the risk minimization framework is essentially the expected utility value, which reflects both the user's preferences and the uncertainty of the

query and document models. Such a notion of relevance is clearly more general than the traditional notion of independent topical relevance, since the utility can depend on factors that might affect a user's satisfaction with the system's action. For example, such factors may include a user's perception of redundancy or special characteristics of documents or the collection. This can be seen formally from the dependency of the loss function on variables such as $\mathcal{U}$, $\vec{\mathcal{S}}$, and $\mathcal{C}$.

The traditional notion of independent relevance can be obtained as a special case of this general utility notion by making an independence assumption on the loss function. Under this assumption, the optimal ranking is to rank documents based on their respective expected loss/risk. This expected risk essentially "measures" the relevance status of a document with respect to a query. It is interesting to note that such a measure explicitly captures two different types of uncertainty. First, it is assumed that the "content" or "topic" (represented by a model) underlying a document or query is uncertain; given a document or a query, we can only estimate the model. This uncertainty reflects the system's inability to completely understand the underlying content/topic of a query or document, so it can be called "topic uncertainty." Second, even if we know the true model for the query and the document, the relevance value of the document model with respect to the query model is still uncertain and vague. This uncertainty reflects our incomplete knowledge of the user's true relevance criterion, and can be called "relevance uncertainty." The topic uncertainty is handled through computing an expectation over all possible models, while the relevance uncertainty is resolved through the specification of a concrete loss function.

As we make different assumptions to simplify the computation of the risk minimization formula, we end up resolving this uncertainty in different ways. In the similarity-based model, for example, we resolve the topic uncertainty by choosing the most likely model and relying on a similarity/distance function to measure the relevance uncertainty. The probabilistic relevance model (including the language modeling approach), however, assumes a binary relevance relationship between a query and a document, and addresses the relevance uncertainty and the topic uncertainty within a single probabilistic model. With a binary relevance relationship, a document is either relevant or non-relevant to a query. Thus, the degree of relevance is not modeled.

7 Conclusions

This paper presents a general probabilistic framework for text retrieval based on the framework of Bayesian decision theory. In this framework, queries and documents are modeled using statistical language models, user preferences are modeled through loss functions, and retrieval is cast as a risk minimization problem. This risk minimization framework not only unifies several existing retrieval mod-

els within a single probabilistic framework, but also facilitates the development of new approaches to text retrieval through the use of statistical language models. We have discussed how special cases of the framework cover existing retrieval models and lead to new models for subtopic retrieval that go beyond independent relevance.

A fundamental difference between the risk minimization framework and previous retrieval frameworks is that the approach presented here treats retrieval as a decision problem, and incorporates statistical language models as major components in the framework. While previous work has treated indexing in a decision-theoretic view, no previous work has given a complete decision-theoretic formal model. The decision space may in principle consist of all the possible actions that the system can take in response to a query. Such a general decision-theoretic view allows retrieval to be modeled as an interactive process that involves cycles of a user's reformulating the query and the system's presenting information. Indeed, one can condition the current retrieval decision on information about the retrieval context, the user, and the interaction history, in order to perform context-sensitive retrieval.

The risk minimization framework makes it possible to systematically and formally study general optimal retrieval strategies. For example, through making different assumptions about the loss function for ranking we have derived an optimal ranking principle, which addresses several limitations of the probability ranking principle. Specifically, when assuming an independent loss function and a sequential browsing model, we can show that the optimal ranking is according to the expected risk of each document, which can be computed independently for each document. An interesting implication is that such a ranking is optimal whether the user has a high-precision or high-recall retrieval preference.

The risk minimization framework incorporates statistical language models systematically in a retrieval framework. As a result, the retrieval parameters are usually introduced as part of a statistical language model. This makes it possible to exploit statistical estimation methods to improve retrieval performance and set retrieval parameters automatically as demonstrated in (Zhai and Lafferty, 2001, 2002). Due to its generality in formalizing retrieval tasks, the risk minimization retrieval framework further allows for incorporating user factors beyond the traditional notion of topical relevance. We presented language models and dependent loss functions that lead to non-traditional ranking models for the subtopic retrieval task. Preliminary exploration of these non-traditional retrieval models has shown promising results, demonstrating that the risk minimization framework facilitates modeling non-traditional retrieval problems (Zhai, 2002; Zhai et al., 2003).

The special cases discussed in this paper represent only a small step toward exploring the full potential of the risk minimization framework, and interesting future research directions remain. For example, it is possible to further exploit the framework to study automatic parameter setting, document structure analysis, and non-traditional retrieval tasks such as subtopic retrieval. In a real retrieval situation, the

goal of satisfying a user's information need is often accomplished through a series of interactions between the user and the retrieval system. With the risk minimization framework, one can formally incorporate these variables and derive personalized and context-sensitive interactive retrieval models. An interesting direction would be to extend the risk minimization framework to formalize an interactive retrieval process, optimizing the utility over a sequence of retrieval interactions.

Acknowledgments

We thank Jamie Callan, Jaime Carbonell, David A. Evans, W. Bruce Croft, Stephen Robertson, William W. Cohen, Rong Jin, Xiaojin Zhu, and several anonymous reviewers for helpful comments on this work. This research was sponsored in part by the Advanced Research and Development Activity in Information Technology (ARDA) under its Statistical Language Modeling for Information Retrieval Research Program, contract MDA904-00-C-2106.

References

Allan, J., Gupta, R., Khandelwal, V., 2001. Temporal summaries of news topics. In: Proceedings of SIGIR 2001. pp. 10–18.

Berger, A., Lafferty, J., 1999. Information retrieval as statistical translation. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 222–229.

Blei, D., Ng, A., Jordan, M., January 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Bollmann-Sdorra, P., Raghavan, V., 1993. On the delusiveness of adopting a common space for modeling ir objects - are queries documents? Journal of the American Society for Information Science 44, 579–587.

Bookstein, A., Swanson, D., 1975. A decision theoretic foundation for indexing. Journal for the American Society for Information Science 26, 45–50.

Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR 1998.

Cooper, W., 1991. Some inconsistencies and misnomers in probabilistic ir. In: Proceedings of SIGIR'91. pp. 57–61.

Cooper, W. S., 1994. The formalism of probability theory in ir: A foundation for an encumbrance? In: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 242–247.

Cooper, W. S., Maron, M. E., January 1978. Foundations of probabilistic and utility-theoretic indexing. Journal of the ACM 25 (1), 67–80.

Croft, W. B., Nov 1981. Document representation in probabilistic models of information retrieval. Journal of American Society for Information Science, 451–457.

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

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of American Society for Information Science 41, 391–407.

Dominich, S., 2000. A unified mathematical definition of classical information retrieval. Journal of the American Soceity for Information Science 51 (6), 614–624.

Dominich, S., 2001. Mathematical foundations of Information Retrieval. Kluwer Academic Publisher.

Dominich, S., 2002. Paradox-free formal foundation of vector-space model. In: Proceedings of the ACM SIGIR 2002 Workshop on Mathematical/Formal Methods in Information Retrieval. pp. 43–48.

Evans, D. A., Zhai, C., 1996. Noun-phrase analysis in unrestricted text for information retrieval. In: Proceedings of ACL 1996. pp. 17–24.

Fox, E., 1983. Expending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types. Ph.D. thesis, Cornell University.

Fuhr, N., 1992. Probabilistic models in information retrieval. The Computer Journal 35 (3), 243–255.

Fuhr, N., May 31 – June 1 2001. Language models and uncertain inference in information retrieval. In: Proceedings of the Language Modeling and IR workshop. Extended abstract.

Fuhr, N., Buckley, C., 1991. A probabilistic learning approach for document indexing. ACM Transactions on Information Systems 9 (3), 223–248.

Fung, R., Favero, B. D., 1995. Applying Bayesian networks to information retrieval. Communications of the ACM 38 (3), 42–48.

Gey, F., 1994. Inferring probability of relevance using the methodof logistic regression. In: Proceedings of ACM SIGIR'94. pp. 222–231.

Harter, S. P., July-August 1975. A probabilistic approach to automatic keyword indexing (part i & ii). Journal of the American Society for Information Science 26, 197–206 (Part I), 280–289 (Part II).

Hiemstra, D., Kraaij, W., 1998. Twenty-one at TREC-7: Ad-hoc and cross-language track. In: Proc. of Seventh Text REtrieval Conference (TREC-7).

Kalt, T., 1996. A new probabilistic model of text classification and retrieval. Tech. Rep. 78, CIIR, Univ. of Massachusetts.

Kwok, K. L., July 1995. A network approach to probabilistic information retrieval. ACM Transactions on Office Information System 13, 324–353.

Lafferty, J., Zhai, C., Sept 2001. Document language models, query models, and risk minimization for information retrieval. In: Proceedings of SIGIR'2001. pp. 111–119.

Lafferty, J., Zhai, C., 2003. Probabilistic relevance models based on document and query generation. In: Croft, W. B., Lafferty, J. (Eds.), Language Modeling and Information Retrieval. Kluwer Series on Information Retrieval.

Lavrenko, V., Croft, B., Sept 2001. Relevance-based language models. In: Proceedings of SIGIR'2001. pp. 120–127.

Lewis, D. D., 1992. Representation and learning in information retrieval. Tech. Rep. 91-93, Univ. of Massachusetts.

Lewis, D. D., 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In: European Conference on Machine Learning. pp. 4–15.

Maron, M. E., Kuhns, J. L., 1960. On relevance, probabilistic indexing and information retrieval. Journal of the ACM 7, 216–244.

McCallum, A., Nigam, K., 1998. A comparison of event models for nave bayes text classification. In: AAAI-1998 Learning for Text Categorization Workshop. pp. 41–48.

Miller, D. H., Leek, T., Schwartz, R., 1999. A hidden markov model information retrieval system. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 214–221.

Mittendorf, E., Schauble, P., 1994. Document and passage retrieval based on hidden markov models. In: Proceedings of SIGIR'94. pp. 318–327.

Over, P., 1998. TREC-6 interactive track report. In: Voorhees, E., Harman, D. (Eds.), The Sixth Text REtrieval Conference (TREC-6). pp. 73–82, nIST Special Publication 500-240.

Ponte, J., Croft, W. B., 1998. A language modeling approach to information retrieval. In: Proceedings of the ACM SIGIR. pp. 275–281.

Ribeiro, B. A. N., Muntz, R., 1996. A belief network model for ir. In: Proceedings of SIGIR'96. pp. 253–260.

Ribeiro-Neto, B., Silva, I., Muntz, R., 2000. Bayesian network models for information retrieval. In: Crestani, F., Pasi, G. (Eds.), Soft Computing in Information Retrieval: Techniques and Applications. Springer Verlag, pp. 259–291.

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., 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of SIGIR'94. pp. 232–241.

Robertson, S., Walker, S., 1997. On relevance weights with little relevance information. In: Proceedings of SIGIR'97. pp. 16–24.

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

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.

Robertson, S. E., van Rijsbergen, C. J., F.Porter, M., 1981. Probabilistic models of indexing and searching. In: et al., O. R. N. (Ed.), Information Retrieval Research. Butterworths, pp. 35–56.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523.

Salton, G., McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

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

Salton, G., Yang, C. S., Yu, C. T., Jan-Feb 1975b. A theory of term importance in automatic text analysis. Journal of the American Society for Information Science 26 (1), 33–44.

Singhal, A., Buckley, C., Mitra, M., 1996. Pivoted document length normalization. In: Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 21–29.

Song, F., Croft, B., 1999. A general language model for information retrieval. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 279–280.

Sparck Jones, K., Walker, S., Robertson, S. E., 2000. A probabilistic model of information retrieval: development and comparative experiments - part 1 and part 2. Information Processing and Management 36 (6), 779–808 and 809–840.

Strzalkowski, T., 1997. NLP track at TREC-5. In: Harman, D. (Ed.), Proceedings of the Fifth Text REtrieval Conference (TREC-5).

Turtle, H., Croft, W. B., July 1991. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems 9 (3), 187–222.

van Rijbergen, C. J., 1977. A theoretical basis for theuse of co-occurrence data in information retrieval. Journal of Documentation, 106–119.

van Rijsbergen, C. J., 1979. Information Retrieval. Butterworths.

van Rijsbergen, C. J., 1986. A non-classical logic for information retrieval. The Computer Journal 29 (6).

Wong, S. K. M., Bollmann, P., Yao, Y. Y., 1991. Information retrieval based on axiomatic decision theory. International Journal of General Systems 19 (2), 101–117.

Wong, S. K. M., Yao, Y. Y., 1989. A probability distribution model for information retrieval. Information Processing and Management 25 (1), 39–53.

Wong, S. K. M., Yao, Y. Y., 1995. On modeling information retrieval with probabilistic inference. ACM Transactions on Information Systems 13 (1), 69–99.

Zhai, C., March 31 – April 3 1997. Fast statistical parsing of noun phrases for document indexing. In: 5th Conference on Applied Natural Language Processing (ANLP-97). pp. 312–319.

Zhai, C., 2002. Risk minimization and language modeling in text retrieval. Ph.D. thesis, Carnegie Mellon University.

Zhai, C., Cohen, W. W., Lafferty, J., 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: Proceedings of SIGIR 2003. pp. 10–17.

Zhai, C., Lafferty, J., 2001. Model-based feedback in the KL-divergence retrieval model. In: Tenth International Conference on Information and Knowledge Management (CIKM 2001). pp. 403–410.

Zhai, C., Lafferty, J., Aug 2002. Two-stage language models for information retrieval. In: Proceedings of SIGIR'2002. pp. 49–56.