A Risk Minimization Framework for Information Retrieval

ChengXiang Zhai[1] John Lafferty[2]

[1]Department of Computer Science
University of Illinois at Urbana-Champaign

[2]School of Computer Science
Carnegie Mellon University

Abstract. This paper presents a novel 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 (i.e., probabilistic models of text), 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 the systematic development of new retrieval models. As an example of using the framework to model non-traditional retrieval problems, we derive new 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 go beyond independent topical relevance.

1 Introduction

A large number of different information retrieval models have been proposed and studied over the past decades. However, despite all the progress, no single unified retrieval model has proven to be most effective, and there are still several major challenges. First, theoretical guidance and formal principles have perhaps rarely led directly to good performance; 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. Second, most existing retrieval models are developed based on the assumption of independent topical relevance, which rarely holds in real applications. It is unclear how we may develop models that can relax such an assumption.

In this paper we present a novel probabilistic information retrieval framework that addresses these challenges. The basic idea of the new 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 the standard retrieval problem can be regarded as a decision problem where the decision involves choosing the best ranking. We formalize this decision-theoretic view of retrieval within the framework of Bayesian decision theory. In particular, we treat both 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 with respect to 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, or the expected loss associated with the chosen action conditioned on the observed query and documents.

This new framework unifies several existing retrieval models, including the recently proposed language modeling approach, within one general probabilistic framework, and provides guidance on how we may 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 important advantages over traditional retrieval frameworks. First, it systematically incorporates statistical language models as components. Statistical language models provide a principled way to model text documents and queries, making it possible to set retrieval parameters through statistical estimation methods. Second, 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 can derive an optimal ranking principle, which is similar to the probability ranking principle, but addresses several limitations of it. Finally, the risk minimization framework goes beyond the traditional notion of independent, topical relevance. As we will show later, it is possible to derive 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 point out how the risk minimization is related to them. In Section 3, we present the basic idea and setup of the risk minimization framework. In Section 3.2.1 and Section 3.2.2, we derive several special cases of the framework, and demonstrate how it can cover existing retrieval models and how it can facilitate development of new retrieval models, including those appropriate for the non-traditional subtopic retrieval task. Finally, we summarize the framework in Section 5 and Section 6.

2 Existing Retrieval Models

Over the decades, many different retrieval models have been proposed, studied, and tested. Their mathematical basis spans a large spectrum, including algebra, logic, probability and statistics. It is impractical to provide a complete survey of all the existing retrieval models in this paper, but we can roughly classify the existing models into three major categories, depending on how they define/measure relevance (Dominich, 2001). In the first category, relevance is assumed to be correlated with the similarity between a query and a document. In the second 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. We now discuss the three categories in details.

2.1 Similarity-based Models

In a similarity-based retrieval model, 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 components: (1) a term vector representation of query; (2) a term vector representation of document; (3) a similarity/distance measure of the document vector and the 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 can all be arbitrary in principle (Dominich, 2002).

The flexibility of vector space model makes it easy to incorporate different indexing models. For example, the 2-Poisson probabilistic indexing model can be used to select indexing terms and/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, and thus 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 for the vector space model is that it provides no formal framework for the representation, making the study of representation inherently separated from the relevance estimation. The separation of the relevance function from the weighting of terms has the advantage of being flexible, but makes it very difficult to study the interaction of 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 so far largely heuristic. The two central problems in document and query representation are the extraction of indexing terms/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), though some more recent evaluation has shown more promising improvement on average through using 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 was 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). The discrimination value of an indexing term is the increase or the decrease in the mean inter-document distance caused by adding the indexing term to the term space for text representation. They 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, it is not modeling relevance, but rather, replies on a given similarity measure. Second, it is only helpful for selecting indexing terms, but not very much for the weighting of terms.

The risk minimization framework would suggest 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 the more principled estimation of parameters. The traditional vector space models can be regarded as special cases of this more general similarity model when the parameters are set heuristically.

2.2 Probabilistic Relevance Models

In a probabilistic relevance model, we are 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 a system can never be sure about the true relevance status of a document, so it has to rely on a probabilistic relevance model to estimate it.

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 $\mathbf{r}$ ("relevant") and $\overline{r}$ ("not relevant"). The task is to estimate the probability of relevance, i.e., $p(R = \mathbf{r} \mid D, Q)$. Depending on how this probability is estimated, there are several special cases of this general probabilistic relevance model.

First, $p(R = \mathbf{r} \mid D, Q)$ 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 (Fox, 1983), where features such as term frequency, authorship, and co-citation were combined using linear regression. Fuhr and Buckley (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 principled way of exploring heuristic features and ideas. 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, because regression models are based on heuristic features in the first place, much empirical experimentation would be needed in order to find a set of good features. A regression model thus provides only limited guidance for extending a retrieval model.

Alternatively, $p(R = \mathbf{r} \mid D, Q)$ can be estimated *indirectly* using a generative model, and documents can be ranked according to the following log-odds ratio (Lafferty and Zhai, 2003):

$$\log \frac{p(\mathbf{r} \mid D, Q)}{p(\overline{r} \mid D, Q)} \quad = \quad \log \frac{p(D, Q \mid r)\, p(r)}{p(D, Q \mid \overline{r})\, p(\overline{r})}\,. \tag{1}$$

There are two different ways to factor the conditional probability $p(D, Q \mid R)$ corresponding to "document-generation" and "query-generation," which, as discussed in (Lafferty and Zhai, 2003), would lead to models that have important differences from an estimation perspective, as they involve different parameters for estimation.

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 use of 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 (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 the term frequency into the model, though not directly proposed, but implied by much work in text categorization, is by regarding 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 hard to estimate all the parameters in practice. In (Lafferty and Zhai, 2003) it is argued that the recently proposed language modeling approach to retrieval is also a special probabilistic relevance model when query-generation is used to decompose the generative model. This work provides a relevance-based justification for this new family of probabilistic models based on statistical language modeling.

The language modeling approach was first introduced by Ponte and Croft in (Ponte and Croft, 1998) and independently explored or later explored in (Hiemstra and Kraaij, 1998; Miller et al., 1999; Berger and Lafferty, 1999; Song and Croft, 1999), among others. 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).

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

The language modeling approach has two important contributions. First, it introduces a new 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].

Instead of imposing a strict document-generation or query-generation decomposition of $p(D, Q \mid R)$, one can also "generate" a document-query pair simultaneously. Mittendorf & Schauble (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 in terms of 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 passage and the other 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.

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. In the risk minimization framework, we also maintain a separate generative model for queries and documents respectively, thus support both document generation and query generation in some sense.

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/proving the query from the document. Depending on how one defines what it means by "proving a query from a document," different inference models are possible.

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 proved from the document. Boolean retrieval model can be regarded as a simple case of this model. To cope with the inherent uncertainty in 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. In (Wong and Yao, 1995), Wong and Yao extended the probabilistic inference model and showed that the general probabilistic inference model actually subsumes several other TR models such as Boolean, vector space, and the classic probabilistic models. Fuhr shows that some particular form of the language modeling approach can also be derived as a special case of the general probabilistic inference model (Fuhr, 2001). Nie recently shows that query translation in cross-language information retrieval is a special case of query expansion which can be formulated using logical inference (Nie, 2003).

While theoretically interesting, the probabilistic inference models all must rely on further assumptions about the representation of documents and queries in order to obtain an operational retrieval formula. The choice of such representations is in a way outside the model, so there is little guidance on how to choose or how to improve a representation.

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). 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 infor-

[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.

mation need, one can obtain many different existing specific TR models, such as Boolean, extended Boolean, vector space, and conventional probabilistic models. More importantly, it can potentially go beyond the traditional notion of topical relevance. The generality makes it possible to combine multiple evidence, including different formulations of the same query. The query language based directly on the model has been an important and practical contribution to IR technology.

However, despite its generality, the inference network framework says little about how one can further decompose the general probabilistic model. As a result, operationally, one usually has to set probabilities based on heuristics, as done in the Inquery system (Callan et al., 1992).

Kwok's network model may also be considered as performing a probabilistic inference (Kwok, 1995), though it is based on spread activation.

In general, the probabilistic inference models address the issue of relevance in a very general way. In some sense, the lack of a commitment to specific assumptions in these general models has helped to maintain their generality as a retrieval model. But this also deprives them of the "predictive power" as a theory. As a result, they generally provide little guidance on how to refine the general notion of relevance.

The risk minimization framework is also quite general. Indeed, we will be able to show that many existing models are special cases of risk minimization. Furthermore, the framework goes beyond the traditional notion of topical relevance, just like the inference network framework, and it allows for incorporating multiple user factors as retrieval criteria. However, the risk minimization framework is different from the probabilistic inference models and other Bayesian belief network models in that it provides an explicit and direct connection to (query and document) language models. Techniques of language modeling can thus be brought into an operational retrieval model easily. In this sense, it is a much more refined and operational framework than probabilistic inference models.

3 The Risk Minimization Framework

The risk minimization framework was first presented in (Lafferty and Zhai, 2001). Its basic idea is to formally treat information retrieval as a statistical decision problem, so let us first discuss informally how one may view retrieval as a decision problem.

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. The cycle continues like this. At each 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 information available 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 order (like in Boolean retrieval); alternatively, it may decide to select all the documents and present them as a ranked list (like in the vector space model). In general, there could be many choices for the decision space, and we can regard the whole process of information retrieval as consisting of a series of such decision making tasks for the system.

We now formally define this decision problem. First, we formally define what a query is and what a document is. 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 $p(\theta_Q \,|\, \mathcal{U})$. Using this model, a query $\mathbf{q}$ is then generated with probability $p(\mathbf{q} \,|\, \theta_Q)$. 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

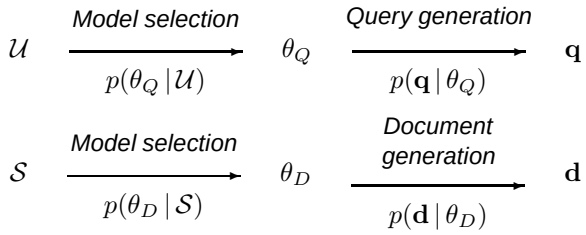

Figure 1. Generative model of query $\mathbf{q}$ and document $\mathbf{d}$.

framework, we will simply refer to $\mathcal{U}$ as a user. Similarly, the source selects a document model θ_D according to a distribution $p(\theta_D \,|\, \mathcal{S})$, and then uses this model to generate a document $\mathbf{d}$ according to $p(\mathbf{d} \,|\, \theta_D)$. Thus, we have Markov chains $\mathcal{U} \rightarrow \theta_Q \rightarrow \mathbf{q}$ and $\mathcal{S} \rightarrow \theta_D \rightarrow \mathbf{d}$. This is illustrated in Figure 1.

Let $\mathcal{C} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N\}$ be a collection of documents obtained from sources $\vec{\mathcal{S}} = (\mathcal{S}_1, ..., \mathcal{S}_N)$. Our observations are thus $\mathcal{U}$, $\mathbf{q}$, $\vec{\mathcal{S}}$, and $\mathcal{C}$. With this setup, we can now formally define retrieval actions. Informally, 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 $\mathbf{q}$ according to some presentation strategy π. Let Π be the set of all possible presentation strategies. We can represent all actions by $\mathcal{A} = \{(D_i, \pi_i)\}$, where $D_i \subseteq \mathcal{C}$ is a subset of $\mathcal{C}$ and $\pi_i \in \Pi$ is some presentation strategy.

In the general framework of Bayesian decision theory, to each such action $a_i = (D_i, \pi_i) \in \mathcal{A}$ there is associated a *loss* $L(a_i, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$, 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 $F(\vec{\mathcal{S}})$. θ_i is the model that generates document $\mathbf{d}_i$.

The *expected risk of action* a_i is given by

$$R(D_i, \pi_i \,|\, \mathcal{U}, \mathbf{q}, \vec{\mathcal{S}}, \mathcal{C}) \;\; = \;\; \int_\Theta L(D_i, \pi_i, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))\, p(\theta \,|\, \mathcal{U}, \mathbf{q}, \vec{\mathcal{S}}, \mathcal{C})\, d\theta$$

where the posterior distribution is given by

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

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

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

That is, to select D^* and present D^* 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. Practically, 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 care about 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.

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

The loss function can encode the user's preferences on the selected subset. Generally, the loss function will have to do with the *relevance* status of the documents selected so that the optimal subset should contain the documents that are most likely 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 have no uncertainty about the query models and document models (e.g., $\theta_Q = \mathbf{q}$ and $\theta_i = \mathbf{d}_i$), and the following loss function is used:

$$
L(D, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) = \sum_{\mathbf{d} \in D} -\delta(\mathbf{d}, \mathbf{q})
$$

where $\delta(\mathbf{d}, \mathbf{q}) = 1$ if and only if document $\mathbf{d}$ satisfies the Boolean query $\mathbf{q}$; otherwise $\delta(\mathbf{d}, \mathbf{q}) = -1$. This loss function is actually quite general, in the sense that if we allow $\delta(\mathbf{d}, \mathbf{q})$ to be any *deterministic* retrieval rule applied to query $\mathbf{q}$ and document $\mathbf{d}$, such that $\delta(\mathbf{d}, \mathbf{q}) > 0$ if $\mathbf{d}$ is relevant to $\mathbf{q}$, otherwise $\delta(\mathbf{d}, \mathbf{q}) < 0$, then the loss function would 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 $a_i = (D_i, \pi_i)$, where π_i is a complete ordering on D_i[3]. Taking action a_i would then mean presenting the selected documents in D one by one in the order given by π_i. This means that we can denote an action by a *sequence* of documents. So we will write $a_i = (d_{\pi_i^1}, d_{\pi_i^2}, ..., d_{\pi_i^k})$, where π_i^j is the index of the document ranked at the j-th rank according to the permutation mapping π_i.

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

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

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

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."

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

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 given by $F(\mathcal{U})$.

$$L(\pi, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) = \sum_{i=1}^{N} s_i l(\pi(1:i), \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$$

where $l(\pi(1:i), \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$ is the actual loss that would be incurred if the user actually views the first i documents according to π. Note that $L(\pi, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$ and $l(\pi, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$ 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:

$$l(\pi(1:i), \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) = \sum_{j=1}^{i} l(d_{\pi^j} | d_{\pi^1}, ..., d_{\pi^{j-1}}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$$

where $l(d_{\pi^j} | d_{\pi^1}, ..., d_{\pi^{j-1}}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}}))$ 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 these together, we have

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

Define the following *conditional* risk

$$\begin{aligned}
r(\mathbf{d}_k | \mathbf{d}_1, ..., \mathbf{d}_{k-1}, \mathbf{q}, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) &\stackrel{\text{def}}{=} \int_{\Theta} l(\mathbf{d}_k | \mathbf{d}_1, ..., \mathbf{d}_{k-1}, \theta, F(\mathcal{U}), F(\vec{\mathcal{S}})) \\
&\qquad \times \; p(\theta | \mathbf{q}, \mathcal{C}, \mathcal{U}, \vec{\mathcal{S}}) d\theta
\end{aligned}$$

which can be interpreted as the expected risk of the user's viewing document $\mathbf{d}_k$ given that $\mathbf{d}_1, ..., \mathbf{d}_{k-1}$ have already been previously viewed. We can write

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

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 would be somehow "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 decomposition would not be appropriate.

We now discuss two general cases of the loss function.

3.2.1 Independent Loss Functions

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

$$l(d_{\pi^j}|d_{\pi^1},...,d_{\pi^{j-1}},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}})) = l(d_{\pi^j},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}}))$$

which means

$$l(\pi(1:i),\theta,F(\mathcal{U}),F(\vec{\mathcal{S}})) = \sum_{j=1}^{i} l(d_{\pi^j},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}}))$$

In this case, the expected risk for ranking π is

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

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:

$$r(\mathbf{d}|\mathbf{q},\mathcal{C},\mathcal{U},\vec{\mathcal{S}}) = \int_{\Theta} l(\mathbf{d},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}}))p(\theta|\mathbf{q},\mathcal{C},\mathcal{U},\vec{\mathcal{S}})d\theta \quad (2)$$

This is equivalent to the situation when we assume each possible action is to present a *single* document. The loss function $l(\mathbf{d},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}}))$ can be interpreted as the loss associated with presenting/viewing document $\mathbf{d}$, or equivalently the expected utility of presenting document $\mathbf{d}$. Equation 2 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 5.

In general, there could be many different ways of specifying the loss function, and a different loss function would lead to a different ranking function. It has been shown in (Lafferty and Zhai, 2001) that with appropriate choices of reasonable loss functions, many existing rank-based retrieval models can be derived as in the risk minimization framework, including the vector space model, classific probabilistic retrieval model, and the recently proposed language modeling approach, which is actually based on the same notion of probability of relevance as the classic probabilistic retrieval model (Lafferty and Zhai, 2003). It has also been shown in (Zhai and Lafferty, 2002; Zhai, 2002) that *new* effective retrieval models, particularly those using statistical language models, can be systematically developed using the risk minimization framework.

3.2.2 Dependent Loss Functions

We have demonstrated that how the risk minimization framework can recover existing retrieval models and 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 is 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 2) 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 $(d_{\pi(1:i)^1}, ..., d_{\pi(1:i)^i}, d_k)$. Then, the increase of risk for picking d_k at rank $i+1$ is

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

To extend $\pi(1:i)$, we should choose

$$\begin{aligned}k^* &= \arg\min_k \delta(k|\pi(1:i)) \\ &= \arg\min_k r(d_k|d_{\pi(1:i)^1}, ..., d_{\pi(1:i)^i}, \mathbf{q},\mathcal{C},\mathcal{U},\vec{\mathcal{S}})\end{aligned}$$

Thus, at each step we just need to evaluate

$$\delta^{'}(k|\pi(1:i)) = r(d_k|d_{\pi(1:i)^1}, ..., d_{\pi(1:i)^i}, \mathbf{q},\mathcal{C},\mathcal{U},\vec{\mathcal{S}}) \tag{3}$$

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

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 this 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.

4 Models for Subtopic Retrieval

4.1 The problem of subtopic retrieval

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

How can we formally define a retrieval model for such a subtopic retrieval problem? Clearly, this would require 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, or subtopics. To model the subtopic retrieval task in the risk minimization framework, we would need a dependent loss function. One possible loss function is the Maximal Marginal Relevance (MMR) loss function, in which we encode the preference for retrieving documents that are both *topically relevant* and *novel* (Carbonell and Goldstein, 1998). Essentially, we want to retrieve relevant documents, and at the same time, minimize the chance for a user to see redundant documents as the user 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 may be more likely. We now discuss this type of loss function in detail.

4.2 Maximal Marginal Relevance (MMR) Loss Functions

The idea of Maximal Marginal Relevance (MMR) ranking is first proposed and formalized in (Carbonell and Goldstein, 1998). It is based on the assumption that we need to consider not only the relevance value, but also the novelty (or equivalently, redundancy) in the presented documents when ranking a set of 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 $l(d_k|d_1,...,d_{k-1},\theta,F(\mathcal{U}),F(\vec{\mathcal{S}}))$ that "balances" the *relevance* value and the *redundancy* value of a document.

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

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

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.,

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

An MMR loss function would generally be a combination of relevance measure and novelty measure to reflect our desire of retrieving a document that is both relevant and novel. Technically, there could be many different ways to specify such a loss function. Indeed, deriving a well-motivated one is still an open research question (Zhai, 2002).

Suppose we make the assumption that a relevance score and a novelty score can be computed *independently*. Then, we can define our loss function as a direct combination of the two scores. Formally, 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 be defined as a combination of the two scoring functions as follows.

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

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

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

One possible such combination is a linear interpolation of S_R and S_N given by

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

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 function S_R and that of S_N are comparable (e.g., when both S_R and S_N are KL-divergence or other comparable functions).

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

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

where c_1, c_2, c_3, and c_4 are cost constants; $p(Rel|\mathbf{d})$ is the probability that document $\mathbf{d}$ is relevant; and $p(New|\mathbf{d})$ is the probability that $\mathbf{d}$ is new with respect to documents $\mathbf{d}_1,...,\mathbf{d}_{k-1}$.

Since whether a non-relevant document carries any new information is not interesting to the user, we could reasonably assume that $c_3 = c_4$. Furthermore, we can also reasonably assume that there is no cost if the document is both relevant and (100%) new, i.e., $c_1 = 0$. Under these two assumptions, we have

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

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 $d_1, ..., d_{k-1}$. Intuitively, c_2 is the cost of seeing a *relevant*, but *redundant* document, whereas c_3 the cost of seeing a *non-relevant* document. Clearly, when $c_2 = 0$, i.e., the user does not care redundancy, the loss function would be essentially based on the probability of relevance, just like what we would expect. Below we assume that $c_2 > 0$, which allows us to re-write the loss function in the following equivalent form for the purpose of ranking documents:

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

Note that a higher $p(New \,|\, \mathbf{d})$ always helps reduce the loss, and when $\frac{c_3}{c_2} \geq 1$, a higher $p(Rel \,|\, \mathbf{d})$ also means a smaller loss. However, the amount of loss reduction is affected by the cost ratio $\frac{c_3}{c_2}$. This ratio 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., $c_3 >> c_2$, the influence of $p(New \,|\, \mathbf{d})$ could be negligible. This means that when the user has low tolerance for any non-relevant document, our 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 $p(Rel|d)p(New|d)$, which is essentially the scoring formula for generating temporal summaries proposed in (Allan et al., 2001), where $p(Rel|d)$ is referred as $p(Useful|d)$. In practice, there would be a compromise between retrieving documents with new content and avoid retrieving non-relevant documents.

In (Zhai, 2002; Zhai et al., 2003), this loss function is explored with $p(Rel|\mathbf{d})$ being assumed to be proportional to $p(\mathbf{q}|\mathbf{d})$ and $p(New|\mathbf{d})$ being estimated with a mixture language model.

A common deficiency in the way we combine the relevance score and the novelty score in our MMR loss function is the assumption of *independent* measurement of relevance and novelty. In other words, we do not have a direct measure of relevance of the new information contained in a new document. Thus a document formed by concatenating a seen (thus redundant) relevant document with a lot of new, but non-relevant information may be ranked high, 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).

It is important to note that there are other dependent loss functions that may be more appropriate for the subtopic retrieval problem; indeed MMR loss functions are not optimizing the subtopic coverage *directly*. Another interesting type of loss function is the Maximal Diverse Relevance (MDR) loss function, in which we encode the preference for retrieving documents that best supplement the previously retrieved documents in terms of covering all the subtopics. This means we would need to model both *topical relevance* and *subtopics* of documents. Some preliminary exploration of the MDR loss functions has been reported in (Zhai, 2002) where the Probabilistic Latent Semantic Indexing (PLSI) model (Hofmann, 1999) and the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) have been used for subtopic modeling.

5 Discussion

5.1 A Decision-Theoretic View of Retrieval

Treating retrieval from a decision-theoretic view is not new. In the 1970's, people 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 these early works was limited to a binary decision regarding whether to retrieve a document or regarding whether to assign an index term to a document, and none of the work gave

a *complete* decision-theoretic formal model for retrieval. The treatment of retrieval as a decision problem was also discussed in (Wong et al., 1991; Dominich, 2001).

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 any existing decision-theoretic treatment of retrieval. 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. We believe that this is the first time a user variable ($\mathcal{U}$) and a document source variable ($\mathcal{S}$) have been explicitly and formally introduced in an *operational* retrieval model. Strictly speaking, $\mathcal{U}$ actually represents a user with a certain information need, thus when the same user enters another query, it should be treated as a different $\mathcal{U}$ (Robertson, 2002). However, it is not hard to imagine that we could factor out the real user and the retrieval context by using two separate variables. Also, we have introduced a separate source variable $\mathcal{S}$ for each document, as if they are all independent. This is, again, to simplify the presentation of the framework; we can easily put constraints on these source variables, e.g., requiring all of them to be identical. The explicit introduction of $\mathcal{U}$ and $\mathcal{S}$ makes it possible for us to consider any interesting user factors and document source factors when specifying the loss function. For example, the high-precision versus high-recall preferences can be encoded by assuming a different stopping probability distribution. Interestingly, as shown formally in this chapter, when we assume an independent loss function or use a greedy algorithm to approximate the optimal ranking based on a sequentially additive loss function, the optimal solution does *not* depend on the stopping probability distribution! The redundancy among documents can be captured using a dependent loss function. Other factors such as readability of documents could also be incorporated as long as we have a model for readability.

Another major 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), takes the 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.

The decision-theoretic view of retrieval makes it possible to model an interactive retrieval process as a sequential decision process, where the user variable $\mathcal{U}$ changes over time. Actually, if we allow the system to accept any user response, rather than just a text query, as input, then we are really going beyond retrieval toward a more general (interactive) information access system.

5.2 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.

From 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:

- Independent loss function: We assume that the loss associated with a user's viewing one document does not depend on any other documents that the user may have seen.
- Sequential browsing: We assume that, when presented with a ranked list of documents, a user would browse through the list sequentially according to the ranking.

It is interesting to note the difference and 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 (Robertson, 2002), 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. This assumption that a greedy algorithm is used to construct the optimal ranking is implicit in (Robertson, 1977), since the decision problem involves whether to retrieve one 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, the PRP 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 more formalized way to go beyond the PRP.

Indeed, 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 ($\mathbf{q}$), 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.

5.3 The Notion of Relevance

The risk minimization framework was originally motivated by the need for a general ranking function that allows us 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 operationality 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 *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 all the 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 independent 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 uncertainties. 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 approximation assumptions to simplify the computation of the risk minimization formula, we end up resolving these uncertainties in different ways. In the general similarity-based model, for example, we resolve the topic uncertainty by picking the most likely model and rely 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 one single probabilistic model. With a binary relevance relationship, a document is either relevant or non-relevant to a query, nothing in between, i.e., the different degree of relevance is not modeled; this is different from the similarity-based model.

5.4 Statistical Language Models for Text Retrieval

The use of language models in the risk minimization framework makes the framework quite different from other general retrieval frameworks such as Kraft's fuzzy models (Kraft et al., 1998), Situation models (Huibers and Bruza, 1996), and "Axiomatic models" (Dominich, 2000). In particular, it makes the framework more *operational*. Indeed, an operational document ranking formula can always be derived by specifying three components: (1) The query model $p(\mathbf{q} \mid \theta_Q)$ and $p(\theta_Q \mid \mathcal{U})$; (2) The document model $p(\mathbf{d} \mid \theta_D)$ and $p(\theta_D \mid \mathcal{S})$; (3) The loss function. A different specification of these components leads to a different operational model.

It is thus clear that if there is any parameter involved in the retrieval formula derived from the risk minimization framework, then it would be from either the loss function or the language models for documents and queries. Parameters associated with the loss function generally represent a user's retrieval preferences, and thus should be set by the user in some meaningful way. For example, the level of redundancy tolerance could be such a parameter, and it must be set by a user, since different users may have different preferences; a high-recall preference may imply more tolerance of redundancy. On the other hand, parameters associated with the language models, in principle, can be estimated automatically. For example, in the following chapters, we will see parameters that control the smoothing of language models. Because such parameters are involved in statistical language models, it is possible to exploit statistical estimation methods to estimate the values of these parameters, thus providing a principled way for setting retrieval parameters.

Being able to estimate retrieval parameters is a major advantage of using language models for information retrieval. For example, the two-stage language model has been show to achieve excellent retrieval performance through the completely automatic setting of retrieval/smoothing parameters (Zhai and Lafferty, 2002).

Another advantage of using language models is that we can expect to achieve better retrieval performance

through the more accurate estimation of a language model or through the use of a more reasonable language model. Thus, we will have more guidance on how to improve a retrieval model than in a traditional model. This will be demonstrated by another new retrieval model, in which feedback documents are exploited to improve the estimation of the query language model. We will show that the improved query model does indeed lead to improved retrieval performance in general.

Finally, language models are also useful for modeling the sub-topic structure of a document and the redundancy between documents. These will also be explored in the thesis as a way to achieve a non-traditional ranking of documents, e.g., to minimize redundancy or maximize sub-topic coverage.

6 Conclusions and Future Work

This paper presents a new general probabilistic framework for text retrieval based on 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 models within one general probabilistic framework, but also facilitates the development of new principled approaches to text retrieval through the use of statistical language models. We have discussed how we may derive many interesting special cases of the framework that both cover known existing retrieval models and lead to new models for subtopic retrieval that go beyond independent relevance.

One fundamental difference between the risk minimization framework and any existing retrieval framework is that the risk minimization framework treats the entire retrieval problem as a *decision* problem, and incorporates *statistical language models* as major components in the framework. While previous work has also treated retrieval from a decision-theoretic view, no previous work has given a *complete* decision-theoretic formal model for retrieval. The risk minimization framework is thus the first complete formal treatment of retrieval in statistical decision theory. This is also the first time a user variable ($\mathcal{U}$) and a document source variable ($\mathcal{S}$) have been explicitly and formally introduced in an *operational* retrieval model. The decision space in the risk minimization framework, in principle, may consist of all the possible actions that the system can take in response to a query, which allows us to treat the retrieval problem in the most general way. Such a general decision-theoretic view explicitly suggests that retrieval can be modeled as an *interactive* process that involves cycles of a user's reformulating the query and the system's presenting information. Indeed, with the risk minimization framework, we can condition the current retrieval decision on all the information about the retrieval context, the user, and the interaction history, to perform context-sensitive retrieval. In contrast, the traditional retrieval models are quite restricted due to their reliance on unrealistic simplification assumptions about relevance (e.g., the independent relevance assumption). They are generally inadequate for handling user factors such as redundancy tolerance and readability, and cannot model an interactive retrieval process without relying on heuristics.

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 to rank documents 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 general incorporation of statistical language models in a retrieval framework is another important contribution of the risk minimization framework. In a traditional model, the parameters are often empirically motivated, so heavy empirical tuning of parameters is always necessary to achieve good retrieval performance. In contrast, the retrieval parameters in the risk minimization framework are generally 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; Zhai and Lafferty, 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. Traditionally, it has been hard to formally model such factors as redundancy and sub-topics *within* a retrieval model, though a general linear

combination of relevance measure and novelty measure has been given in (Carbonell and Goldstein, 1998). 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 (Zhai, 2002; Zhai et al., 2003), demonstrating that the risk minimization framework can be exploited to model such a non-traditional retrieval problem as subtopic retrieval.

The risk minimization framework opens up many new possibilities for developing principled approaches to text retrieval, and serves as a general framework for applying statistical language models to text retrieval. The special cases discussed in this paper represent only a small step in exploring the full potential of the risk minimization framework. There are many interesting future research directions. Naturally, 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, we can formally incorporate all these variables and derive personalized and context-sensitive interactive retrieval models. It would be very interesting to extend the risk minimization framework to formalize an interactive retrieval process so as to optimize the *global* and *long term* utility over a sequence of retrieval interactions.

7 Acknowledgements

We thank Rong Jin, Xiaojin Zhu, Jamie Callan, Jaime Carbonell, David A. Evans, W. Bruce Croft, and William W. Cohen for an anonymous reviewer 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., and Khandelwal, V. (2001). Temporal summaries of news topics. In *Proceedings of SIGIR 2001*, pages 10–18.

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

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Callan, J. P., Croft, W., and Harding, S. (1992). The inquery retrieval system. In *Proceedings of the Third International Conference on Database and Expert System Applications*, pages 78–82. Springer-Verlag.

Carbonell, J. and 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*, pages 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*, pages 242–247.

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

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

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and 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 Society for Information Science*, 51:614–625.

Dominich, S. (2001). *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers.

Dominich, S. (2002). Paradox-free formal foundation of vector space model. In *Proceedings of ACM SIGIR 2002 Workshop on Mathematical/Formal Methods in IR*.

Evans, D. A. and Zhai, C. (1996). Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of ACL 1996*.

Fox, E. (1983). *Expending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types*. PhD thesis, Cornell University.

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

Fuhr, N. (2001). Language models and uncertain inference in information retrieval. In *Proceedings of the Language Modeling and IR workshop*. Extended abstract.

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

Fung, R. and 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*, pages 222–231.

Harter, S. P. (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. and Kraaij, W. (1998). Twenty-one at trec-7: Ad-hoc and cross-language track. In *Proc. of Seventh Text REtrieval Conference (TREC-7)*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR 1999*, pages 50–57.

Huibers, T. W. C. and Bruza, P. D. (1996). Situations, a general framework for studying ir. In *Proceedings of the 16th Research Colloquim of the British Computer Society IR Specialist Group*, pages 3–25.

Kalt, T. (1996). A new probabilistic model of text classification and retrieval. Technical Report 78, CIIR, Univ. of Massachusetts.

Kraft, D. H., Bordogna, P., and Pasi, G. (1998). Fuzzy set techniques in information retrieval. In *Handbook of Fuzzy Sets and Possibility Theory. Approximate Reasoning and Fuzzy Information Systems*. Kluwer Academic Publishers.

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

Lafferty, J. and Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'2001*, pages 111–119.

Lafferty, J. and Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In Croft, W. B. and Lafferty, J., editors, *Language Modeling for Information Retrieval*, volume 13. Kluwer Academic Publishers.

Lewis, D. D. (1992). Representation and learning in information retrieval. Technical Report 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*.

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

McCallum, A. and Nigam, K. (1998). A comparison of event models for nave bayes text classification. In *AAAI-1998 Learning for Text Categorization Workshop*, pages 41–48.

Miller, D. H., Leek, T., and 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*, pages 214–221.

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

Nie, J.-Y. (2003). Query expansion and query translation as logical inference. *Journal of the American Society for Information Science*, 54:335–346.

Over, P. (1998). Trec-6 interactive track report. In Voorhees, E. and Harman, D., editors, *The Sixth Text REtrieval Conference (TREC-6)*, pages 73–82. NIST Special Publication 500-240.

Ponte, J. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR*, pages 275–281.

Ribeiro, B. A. N. and Muntz, R. (1996). A belief network model for ir. In *Proceedings of SIGIR'96*, pages 253–260.

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

Robertson, S. (2002). Personal communication.

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

Robertson, S. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR'94*, pages 232–241.

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

Robertson, S. E., Maron, M. E., and 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., and F.Porter, M. (1981). Probabilistic models of indexing and searching. In et al., O. R. N., editor, *Information Retrieval Research*, pages 35–56. Butterworths.

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

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

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

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

Salton, G., Yang, C. S., and Yu, C. T. (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., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29.

Song, F. and 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*, pages 279–280.

Sparck Jones, K., Walker, S., and 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., editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*.

Turtle, H. and Croft, W. B. (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*, pages 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., and Yao, Y. Y. (1991). Information retrieval based on axiomatic decision theory. *General Systems*, 19(2):101–117.

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

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

Zhai, C. (1997). Fast statistical parsing of noun phrases for document indexing. In *5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 312–319.

Zhai, C. (2002). *Risk Minimization and Language Modeling in Text Retrieval*. PhD thesis, Carnegie Mellon University.

Zhai, C., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR 2003*. To appear.

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

Zhai, C. and Lafferty, J. (2002). Two-stage language models for information retrieval. In *Proceedings of SIGIR'2002*, pages 49–56.