A Brief Review of Information Retrieval Modesl

ChengXiang Zhai

February 6, 2007

Abstract

Information retrieval models have been studied for decades, leading to a huge body of literature on the topic. In this paper, we briefly review this body of literature along with a discussion of some recent trends.

1 Introduction

The goal of any information retrieval (IR) system is to identify documents relevant to a user's query. In order to do this, an IR system must assume some specific measure of relevance between a document and a query, i.e., an *operational* definition of a "relevant document" with respect to a query. A fundamental problem in IR research is thus to formalize the concept of relevance; a different formalization of relevance generally leads to a different retrieval model.

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. The existing models can be roughly grouped into three major categories, depending on how they define/measure relevance. 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 relevance uncertainty is modeled by the uncertainty of inferring queries from documents or vice versa. We now discuss the three categories in details.

2 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 a 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 the vector space model, a document and a query are represented as two term vectors in a high-dimensional term space. 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.

Formally, a document d may be represented by a document vector $\vec{d} = (x_1, x_2, ..., x_n)$, where n is the total number of terms and x_i is the weight assigned to term i. Similarly, a query q can be represented by a query vector $\vec{q} = (y_1, y_2, ..., y_n)$. The weight is usually computed based on the so-called TF-IDF weighting, which is a combination of three factors (Singhal, 2001): (1) the local frequency of the term (in a document or query); (2) the global frequency of the term in the whole collection; (3) document length. With the cosine measure, we have the following similarity function of the document and query:

$$sim(d, q) = \frac{\vec{d} \cdot \vec{q}}{\sqrt{\|\vec{d}\|\|\vec{q}\|}}$$

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.

Related to its generality, the vector space model can also be regarded as a procedural model of retrieval, in which the task of retrieval is naturally divided into two separate stages: indexing and search. The indexing stage explicitly has to do with representing the document and the query by the "indexing terms" extracted from the document and the query. The indexing terms are often assigned different weights to indicate their importance in describing a document or a query. The search stage has to do with evaluating the relevance value (i.e., the similarity) between a document vector and a query vector. The flexibility of the 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).

In the vector space model, feedback is typically treated as query vector updating. A well-known approach is the Rocchio method, which simply adds the centroid vector of the relevant documents to the query vector and subtracts from it the centroid vector of the non-relevant documents with appropriate coefficients (Rocchio, 1971). In effect, this leads to an expansion of the original query vector, i.e., additional terms are extracted from the known relevant (and non-relevant) documents, and are added to the original query vector with appropriate weights (Salton and Buckley, 1990).

The extended Boolean (p-norm) model is a heuristic extension of the traditional Boolean model to perform document ranking, but it can also be regarded as a special case of the similarity model (Fox, 1983; Salton et al., 1983). The similarity function has a parameter p that controls the "strictness" of satisfying the constraint of a Boolean query, in such a way that it approaches a strict (conjunctive or disjunctive) Boolean model when p approaches infinity, but softens the conjunctive or disjunctive constraint and behaves more like a regular vector space similarity measure as p becomes smaller. However, the model must rely on some assumptions about the Boolean structure of a query, and has some undesirable mathematical properties (Rousseau, 1990). There has also been little, if any, large-scale evaluation of the model.

The vector space model is by far the most popular retrieval model due to its simplicity and effectiveness. The following is a typical effective weighting formula with pivoted document length normalization taken from (Singhal, 2001):

$$\sum_{t \in Q,D} \frac{1 + \ln(1 + \ln(tf))}{(1-s) + s\frac{dl}{avdl}} \times qtf \times \ln \frac{N+1}{df}$$

where s is an empirical parameter (usually 0.20), and

tf	is the term's frequency in document
qtf	is the term's frequency in query
N	is the total number of documents in the collection
df	is the number of documents that contain the term
dl	is the document length, and
$avdl$	is the average document length.

The main criticism for the vector space model is that it provides no formal framework for the representation, making the study of representation inherently separate 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 semantics 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 has been found to be significantly better than the heuristic TF-IDF term weighting (Salton and Buckley, 1988). To address the variances 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 a 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 discrimination value is not modeling relevance, but rather, relies on a given similarity measure. Second, it is only helpful for selecting indexing terms, but not for the weighting of terms.

A new generation of similarity-based retrieval models (Lafferty and Zhai, 2001a; Zhai and Lafferty, 2001a; Lavrenko, 2004) has been proposed based on the idea of representing documents and queries with statistical language models. In one of the most effective methods, both a document and a query are represented as a unigram language model (i.e., a word distribution), and the similarity (or rather the distance) between a document and a query is then measured based on the Kullback-Leibler divergence of the two language models (Lafferty and Zhai, 2001a; Zhai and Lafferty, 2001a). The use of language models for IR will be further discussed in the next section.

3 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 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 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 the matching of D and Q. Such a regression model was probably 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, lots of 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, $p(R = r \mid D, Q)$ can be estimated *indirectly* using a generative model in the following way (Lafferty and Zhai, 2003):

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

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

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

There are two different ways to factor the conditional probability $p(D, Q \mid R)$, corresponding to "document generation" and "query generation."

With document generation, $p(D, Q \mid R) = p(D \mid Q, R)p(Q \mid R)$, so we end up with the following ranking formula:

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

Essentially, the retrieval problem is formulated as a two-category document classification problem, though we are only interested in ranking the classification likelihood, rather than actually assigning class labels. Operationally, two models are estimated for each query, one modeling relevant documents, the other modeling non-relevant documents. Documents are then ranked according to the posterior probability of relevance.

Most classic probabilistic retrieval models (Robertson and Sparck Jones, 1976; van Rijsbergen, 1979; Fuhr, 1992) are based on document generation. The Binary Independence Retrieval (BIR) model (Robertson

and Sparck Jones, 1976; Fuhr, 1992) is perhaps the most well known classical probabilistic model. The BIR model assumes that terms are independently distributed in each of the two relevance models, so it essentially uses the Naïve Bayes classifier for document ranking (Lewis, 1998). [1] The BIR retrieval formula is the following (Robertson and Sparck Jones, 1976; Lafferty and Zhai, 2003):

$$\log \frac{p(r \mid D, Q)}{p(\overline{r} \mid D, Q)} \stackrel{\text{rank}}{=} \sum_{t \in D \cap t \in Q} \log \frac{p(t \mid Q, r)(1 - p(t \mid Q, \overline{r}))}{(1 - p(t \mid Q, r))p(t \mid Q, \overline{r})}$$

where $\stackrel{\text{rank}}{=}$ means equivalent in terms of being used for ranking documents.

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). The dependency model achieved significant increases in retrieval performance over the independence model. However, the evaluation was only done on very small collections, and the estimation of many more parameters is a problem in practice (Harper and van Rijsbergen, 1978). Croft investigated how the heuristic term significance weight can be incorporated into probabilistic models in a principled way (Croft, 1981). Another effort to improve document representation involves introducing the term frequency directly into the model by using a multiple 2-Poisson mixture representation of documents (Robertson et al., 1981). This model has not shown empirical improvement in retrieval performance directly, but an approximation of the model using a simple TF formula turns out to be quite effective (Robertson and Walker, 1994). The heuristic retrieval formula BM25 has been successfully used in City University's Okapi system and several other TREC systems (Voorhees and Harman, 2001). A different way of introducing the term frequency into the model, not directly proposed but implied by a lot of work in text categorization, is to regard a document as being generated from a unigram language model (Kalt, 1996; McCallum and Nigam, 1998).

With query generation, $p(D, Q \mid R) = p(Q \mid D, R)p(D \mid R)$, so we end up with the following ranking formula:

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

Under the assumption that conditioned on the event $R = \overline{r}$, the document D is independent of the query Q, i.e., $p(D, Q \mid R = \overline{r}) = p(D \mid R = \overline{r})\, p(Q \mid R = \overline{r})$, the formula becomes

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

There are two components in this model. The major component $p(Q \mid D, r)$ can be interpreted as a "relevant query model" conditioned on a document. That is, $p(Q \mid D, r)$ is the probability that a user who likes document D would use Q as a query. The second component $p(r \mid D)$ is a prior that can be used to encode a user's bias on documents.

Models based on query generation have been explored in (Maron and Kuhns, 1960), (Fuhr, 1992) and (Lafferty and Zhai, 2001b). The probabilistic indexing model proposed in (Maron and Kuhns, 1960) is the 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 the user's feedback, not the content of D. The Binary Independence Indexing (BII) model proposed in (Fuhr, 1992) is another

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

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, 2001b), it has been shown 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 in (Ponte and Croft, 1998) and 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 method of language model estimation. Smoothing document language models 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 a new effective probabilistic ranking function based on query generation. While the earlier query generation models have all found estimating the parameters difficult, 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]

While based on the same notion of relevance and probabilistically equivalent, the classic document generation probabilistic models and the language modeling approach 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 whole collection of documents is non-relevant (Croft and Harper, 1979; Robertson and Walker, 1997). Recently, Lavrenko and Croft made progress in estimating the relevance model without relevance judgments by exploiting language modeling techniques (Lavrenko and Croft, 2001). When explicit relevance judgments are available, the classic models, being based on document generation, have the advantage of being able to improve the estimation of the component probabilistic models naturally 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 judged relevant documents. Thus, the directly improved models can *not* be expected to improve our ranking of other un-judged documents. Interestingly,

[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.

such improved models can potentially be beneficial for new queries – a feature unavailable in document generation models.

The difficulty in doing feedback for the query-generation probabilistic models has motivated the development of a more general family of probabilistic similarity models as discussed in the end of Section 2. It can be shown that the Kullback-Leibler (KL) divergence retrieval can cover the simple query generation model as a special case (Lafferty and Zhai, 2001a). Moreover, with the KL-divergence model, feedback can be achieved through improving the estimation of the query language model (Lafferty and Zhai, 2001a) (also called relevance model (Lavrenko and Croft, 2001)). In general, retrieval performance can be improved through improving the estimation of both the query language model (Zhai and Lafferty, 2001a; Lavrenko and Croft, 2001; Shen et al., 2005; Tao and Zhai, 2006) and the document language model (Liu and Croft, 2004; Kurland and Lee, 2004; Tao et al.,).

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 explored a passage-based generative model using the Hidden Markov Model (HMM), which can be regarded as such a case (Mittendorf and Schauble, 1994). In this work, a document query pair is represented as a sequence of symbols, each corresponding to a term in 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 in 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.

Empirically, probabilistic relevance models have shown good performance. Indeed, a simple approximation of the 2-Poisson probabilistic model, which has led to the BM25 retrieval formula used in the Okapi system, has been very effective (Robertson and Walker, 1994; Robertson et al., 1995). The language modeling approaches have also been shown to perform very well (Ponte and Croft, 1998; Hiemstra and Kraaij, 1998; Miller et al., 1999). The BM25 formula is shown below, following the notations used in (Singhal, 2001):

$$\sum_{t \in Q,D} \ln \frac{N - df + 0.5}{df + 0.5} \times \frac{(k_1 + 1)tf}{(k_1(1 - b) + b\frac{dl}{avdl}) + tf} \times \frac{(k_3 + 1)qtf}{k_3 + qtf}$$

where, $k_1 \in [1.0, 2.0]$, b (usually 0.75), and $k_3 \in [0, 1000]$ are parameters, and other variables have the same meaning as in the vector space retrieval formula described in the previous section.

Also, most language model-based retrieval models, especially the KL-divergence retrieval model with Dirichlet prior smoothing (Zhai and Lafferty, 2001b), perform very well empirically.

4 Probabilistic Inference Models

In a probabilistic inference model, the relevance uncertainty 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 which, a document is relevant to a query if and only if the query can be proved from the document. The 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 developed a general probabilistic inference model which subsumes several other retrieval 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).

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 the Bayesian belief network in retrieval are presented in (Fung and Favero, 1995; Ribeiro and Muntz, 1996; Ribeiro-Neto et al., 2000). The inference network model is a much more general formalism than most of the models that we have discussed above. With different ways to realize the probabilistic relationship between the observation of documents and the satisfaction of the user's information need, one can obtain many different existing specific retrieval models, such as Boolean, extended Boolean, vector space, and conventional probabilistic models. More importantly, the inference network model can potentially go beyond the traditional notion of topical relevance; indeed, the goal of inference is a very general one, and at its highest level, the framework is so general that it can accommodate almost any probabilistic model. 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 was 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 retrieval models. But this also deprives them of "predictive power" as a theory. As a result, they generally provide little guidance on how to refine the general notion of relevance.

5 Recent Trends

A large number of different retrieval approaches have been proposed and studied, and a tremendous amount of effort has been devoted to the evaluation of various kinds of approaches, especially in the context of TREC evaluation (Voorhees and Harman, 2001). There has been a lot of progress in both developing a retrieval theory and improving empirical performance. Among all, the following three basic retrieval functions are generally regarded as most effective: (1) pivoted normalization vector space model (Singhal et al., 1996); (2) Okapi/BM25 probabilistic retrieval model (Robertson and Walker, 1994); and (3) Dirichlet prior language model (Zhai and Lafferty, 2001b). When optimized, they generally have similar performance. Moreover, each of them can be further improve through techniques such as pseudo feedback.

A main weakness in all the work is that the integration of theory and practice has so far been quite weak in the sense that theoretical guidance and formal principles have rarely led to good performance *directly*; a

lot of heuristic parameter tuning must be used in order to achieve good performance. Parameter tuning is generally difficult due to the fact that the optimal setting of parameters is often collection/query dependent and the parameters may interact with each other in a complicated way.

To address these long-standing challenges, two lines of work have been done recently: (1) general retrieval frameworks such as risk minimization (Zhai and Lafferty, 2006) and generative relevance (Lavrenko, 2004) have been developed. These statistically well-founded frameworks can provide a roadmap for applying statistical language models to retrieval. (2) a new axiomatic approach to developing retrieval models has been proposed and explored (Fang et al., 2004; Fang and Zhai, 2005; Fang and Zhai, 2006). This new approach attempts to model relevance more directly with term-level heuristic constraints. As a result, it provides several important benefits, including making it possible to analytically predict the performance of a retrieval function without relying on labor-intensive experiments, determining optimal parameter ranges analytically, and providing guidance on developing new retrieval models.

References

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.

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.

Cooper, W. (1991). Some inconsistencies and misnomers in probabilistic IR. In *Proceedings of SIGIR'91*, pages 57–61.

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

Croft, W. B. and 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., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Evans, D. A. and Zhai, C. (1996). Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of ACL 1996*, pages 17–24.

Fang, H., Tao, T., and Zhai, C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fang, H. and Zhai, C. (2005). An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 2005 ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fang, H. and Zhai, C. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 2006 ACM SIGIR Conference on Research and Development in Information Retrieval*.

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*, pages 6–11.

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 method of logistic regression. In *Proceedings of ACM SIGIR'94*, pages 222–231.

Harper, D. J. and van Rijsbergen, C. J. (1978). An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216.

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 *Proceedings of Seventh Text REtrieval Conference (TREC-7)*, pages 227–238.

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

Kurland, O. and Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 194–201. ACM Press.

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. (2001a). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pages 111–119.

Lafferty, J. and Zhai, C. (2001b). Probabilistic IR models based on query and document generation. In *Proceedings of the Language Modeling and IR workshop*, pages 1–5.

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 and Information Retrieval*. Kluwer Academic Publishers.

Lavrenko, V. (2004). *A generative theory of relevance*. PhD thesis, University of Massachusetts, Amherst, MA.

Lavrenko, V. and Croft, B. (2001). Relevance-based language models. In *Proceedings of SIGIR'01*, pages 120–127.

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*, pages 4–15.

Liu, X. and Croft, W. B. (2004). Cluster-based retrieval using language models. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 186–193. ACM Press.

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.

Ponte, J. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR'98*, 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. 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. and Walker, S. (1997). On relevance weights with little relevance information. In *Proceedings of SIGIR'97*, pages 16–24.

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

Robertson, S. E., Walker, S., Jones, S., M.Hancock-Beaulieu, M., and Gatford, M. (1995). Okapi at TREC-3. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 109–126.

Rocchio, J. (1971). Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall Inc.

Rousseau, R. (1990). Extended Boolean retrieval: a heuristic approach. In *Proceedings of SIGIR'90*, pages 495–508.

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 Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 44(4):288–297.

Salton, G., Fox, E., and Wu, H. (1983). Extended boolean information retrieval. *The Communications of the ACM*, 26(1):1022–1036.

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.

Shen, X., Tan, B., and Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In *Proceedings of SIGIR 2005*, pages 43–50.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43.

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)*, pages 97–102.

Tao, T., Wang, X., Mei, Q., and Zhai, C. Language model information retrieval with document expansion. In *Proceedings of HLT/NAACL 2006*, page 407.

Tao, T. and Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo feedback. In *Proceedings of ACM SIGIR 06*.

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).

Voorhees, E. and Harman, D., editors (2001). *Proceedings of Text REtrieval Conference (TREC1-9)*. NIST Special Publications. http://trec.nist.gov/pubs.html.

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. and Lafferty, J. (2001a). 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. (2001b). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, pages 334–342.

Zhai, C. and Lafferty, J. (2006). A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55.