Probabilistic Models of Novel Document Rankings for Faceted Topic Retrieval

Ben Carterette and Praveen Chandar
{carteret,pcr}@udel.edu
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716

ABSTRACT

Traditional models of information retrieval assume documents are independently relevant. But when the goal is retrieving diverse or novel information about a topic, retrieval models need to capture dependencies between documents. Such tasks require alternative evaluation and optimization methods that operate on different types of relevance judgments. We define *faceted topic retrieval* as a particular novelty-driven task with the goal of finding a set of documents that cover the different facets of an information need. A faceted topic retrieval system must be able to cover as many facets as possible with the smallest number of documents. We introduce two novel models for faceted topic retrieval, one based on pruning a set of retrieved documents and one based on retrieving sets of documents through direct optimization of evaluation measures. We compare the performance of our models to MMR and the probabilistic model due to Zhai et al. on a set of 60 topics annotated with facets, showing that our models are competitive.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]

General Terms: Algorithms, Experimentation

Keywords: information retrieval, novelty, diversity, probabilistic models

1. INTRODUCTION

The concept of relevance is probably the most important and most vociferously debated in the field of information retrieval. Many researchers have settled on a so-called "system-based" notion of relevance that is amenable to fast research and development cycles. In this conception, documents may be relevant on binary, graded, or continuous scales, but all documents are judged relevant independently of one another. Two identical documents are both relevant as long as they contain information the user needs. Most evaluation measures, including precision, recall, average precision, and discounted cumulative gain, assume such a definition; the Probability Ranking Principle for optimizing document rankings is also based on independent relevance.

Modeling documents as independently relevant does not necessarily provide the optimal user experience. Certainly, five relevant documents that all contain the same single piece of information are not as useful to a user as one relevant document that contains five separate pieces of information—yet traditional evaluation measures would reward a system that provides the former more than one that provides the latter. *Novelty* and *diversity* tasks attempt to remedy this with new definitions of relevance and new evaluation measures.

We view such tasks as falling on a continuum: at one end, there is diversity through retrieving different results for independent interpretations of a query, in the way that *Michael Jordan* the basketball player and *Michael Jordan* the statistician largely occur in documents independently of one another. At the other, diversity is achieved through retrieving documents that are all relevant to the same interpretation, but cover different facets of the topic. For the Michael Jordan example, this might entail assuming that the basketball player is the correct interpretation, then ranking documents that cover different aspects of his life and career (e.g. his time with the Chicago Bulls, his time with the Washington Wizards, his gambling problems, etc.) with diversity. In between these two endpoints there is a wide variety of tasks, and the precise types of optimization and evaluation needed may vary significantly from task to task.

In this work we attack a task closer to the latter end of the continuum: *faceted topic retrieval.* Our definition of faceted topic retrieval assumes a single "correct" interpretation of a query; within that interpretation there are multiple facets, all of which must be represented in the retrieved documents. These facets are highly correlated, often appearing together in groups in the same documents. The faceted topic retrieval system must be able to find a small set of documents that covers all of the facets: three documents that cover 10 facets will always be preferable to five documents that cover the same 10. Because of the high correlations among facets, some redundancy in the retrieved results is unavoidable. Evaluation and optimization must take care to not penalize redundancy too much.

We propose a novel set-based probabilistic model for faceted topic retrieval. Our model makes no explicit attempt to control redundancy, yet it performs as well as greedy ranking methods that attempt to minimize redundancy such as

CIKM'09, November 2–6, 2009, Hong Kong, China.

MMR and the probabilistic method of Zhai et al. Furthermore, upper bound experiments suggest that our model has greater flexibility and room for improvement.

2. PREVIOUS WORK

The need for diversity in the result sets was addressed by Goffman in 1964 [8]. He stressed that the relevance of a document is dependent on the previous documents retrieved. Several researchers have been working on ways to eliminate the redundancy in the result set and have proposed models for diverse document ranking.

In their work on subtopic retrieval, Zhai et al. claim that there is more than one meaningful interpretation for a given query [16]. They assume that these interpretations indicate the various subtopics for the query. They re-order the results such that some results from each subtopic are accommodated in the top results with some probability. Their methodology involves handling the novelty and redundancy in a result set separately, then combining them in a cost function. This paper also introduces evaluation measures for subtopic retrieval that we will use for our task as well.

The work of Zhai et al. is based on the Maximum Marginal Relevance (MMR) ranking function of Carbonell and Goldstein [4]. The MMR approach aims to reduce the redundancy and achieve diversity in the result set by ranking documents that are relevant to the query but dissimilar to documents ranked above them. Similar work by Chen and Karger aims directly to provide the user with the answer to their interpretation for the query [5]. Their greedy algorithm incorporates negative feedback in order to maximize diversity in the result set by penalizing redundancy.

Clarke et al. note that the evaluation measure acts as an objective function, and claim that it should reflect user requirements [6]. They introduce an evaluation measure based on normalized discounted cumulative gain (nDCG [10]) that rewards novelty and diversity and penalizes redundancy. Since it discounts by rank, it seems to demand a greedy strategy for optimization. We argue that greedy strategies are not necessarily optimal for maximizing diversity.

The objective of our paper is similar to those of the above mentioned work: to provide a diverse ranking mechanism and remove redundancy. But while the previous work concentrated on maximizing diversity and penalizing redundancy in a single optimization step (usually by means of a greedy algorithm), ours will maximize diversity among a set of documents without regard for redundancy. We wish to retrieve a smallest set of documents that cover a given set of facets, and this goal can conflict with the goal of minimizing the redundancy among a set of documents: consider cases in which a particular facet is rare and only occurs in documents along with several much more common facets.

3. FACETED TOPIC RETRIEVAL

Let us define the faceted topic retrieval task in terms of the type of information need the user has and how that need is best satisfied. A faceted topic retrieval information need is one that has a set of answers—facets—that are clearly delineated. Each of those answers may appear in multiple documents, but each answer is fully contained within at least one document (i.e. a user would not have to read two or more documents to understand how some piece of one of them is related to his or her need).

An example faceted topic retrieval information need is:

> Many countries are trying to reduce their dependence on foreign oil. What strategies have countries, organizations, or individuals proposed or implemented to reduce the demand for oil?

The facets of this need include *invest in next generation technologies, increase use of renewable energy sources, invest in renewable energy sources, double ethanol in gas supply, shift to biodiesel, shift to coal*, and more. Note that facets are not limited to any particular part of speech or type of entity; they can be phrases, named entities, places, objects, or a mix of types.

All of the relevant documents must be on the same topic. While there may be room for different interpretations of a short query, the task definition is that the interpretation in the statement of the information need is the "correct" one. A document is relevant to the need if it contains any of the facets (and support for that facet being relevant).

Each document can contain one or more facets, and each facet can be contained in one or more documents. Figure 1 shows how documents and facets can be related in a bipartite graph. Only relevant documents are shown here; quite a few documents have been judged nonrelevant to this need and thus do not contain any facet.

3.1 Relationship to Other Tasks

Despite the similar name, our task is quite different from "faceted search". In faceted search, items are classified into one or more groups called facets, and the user may narrow or expand her search using those facets [7]. In faceted topic retrieval, the task is to retrieve the individual facets of a particular query. The difference may be best expressed by noting that facets in faceted search are defined globally and independent of any query, while the facets in faceted topic retrieval are defined entirely by the information need.

Faceted topic retrieval is also quite different from recent diversity tasks such as those studied by Clarke et al. [6], Agrawal et al. [1], and Radlinski et al. [12]. In those works, diversity is a matter of satisfying varying user needs in a single ranked list for a query. The assumption is that a user is interested in a subset of the relevant material, and different users may be interested in different subsets. In faceted topic retrieval we assume all users are interested in all of the facets, like the standard ad hoc assumption that all users are interested in all of the relevant material.

Our task is more similar to the "list question" task of the TREC Question Answering track, in which a system must find answers to natural language questions such as "list 8 oil producing states in the United States" [15]. For example, one of our queries is *oil producing countries*; the information need is to find countries that produce and export oil, and also to distinguish OPEC nations from non-OPEC nations. The difference is that instead of extracting the facets and presenting them to the user, we present the user with a ranked list of documents as in traditional retrieval. Clearly, then, the documents in that ranked list should contain as many unique facets as possible. Additionally, instead of a natural language question, the query is a short list of keywords, as in the standard ad hoc retrieval task. The query for the example above, for instance, is *reducing dependency on oil*; it does not explicitly ask for a list of proposed strategies. Finally, relevant facets are not necessarily all of

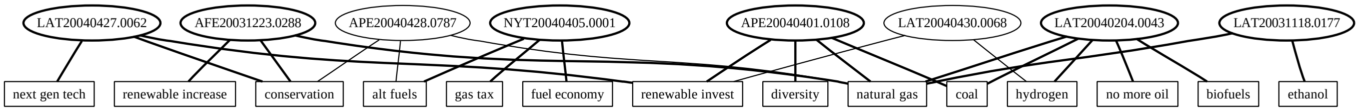

Figure 1: Example document-facet graph. An edge from a document to a facet indicates that the document attests the facet. The bolded document nodes indicate the smallest set needed to cover all of the facets.

the same entity type; people and organizations, people and places, noun phrases and dates, etc. may be mixed together in the facets.

Faceted topic retrieval is most similar to the subtopic retrieval task studied by Zhai et al. [16]. In our view, the difference is primarily of degree: we believe facets will occur together in documents slightly more often than subtopics will. The problems are nevertheless so similar that a human may have difficulty distinguishing between them, and therefore the evaluation methods and models Zhai et al. propose are natural for our task.

3.2 Faceted Topic Retrieval Evaluation

Zhai et al. evaluated subtopic retrieval with measures called *S-recall* and *S-precision*. We use the same measures for our faceted topic retrieval task.

3.2.1 S-Recall

The primary evaluation question for a faceted topic retrieval engine is how many of the facets that are attested in the corpus were retrieved. Given a set of facets and documents judged according to whether they are relevant to the information need and contain each facet, S-recall at rank k may be defined as:

$$\textit{S-rec}@k = \frac{1}{m}\sum_{i=1}^{m} I(F_i \in \{D_1, D_2, ..., D_k\})$$

where m is the number of known facets, D_i is the document retrieved at rank i, and $I(\cdot)$ is the so-called indicator function, which in this case is 1 if F_i occurs in any of the documents ranked $1-k$ and 0 otherwise. This is equivalent to the definition given by Zhai et al. [16], but reformulated in terms of sets of documents.

The maximum value of S-recall at a particular rank k depends on the maximum number of facets that can be found in k documents. For the example in Figure 1, *S-rec*@1 can be at most 5/14 and *S-rec*@2 can be at most 8/14; at least 6 documents are required to achieve *S-rec* = 1. In this example, 6 is the minimum rank at which perfect recall can be achieved, and we will denote S-recall at that rank simply *S-rec*. We argue that the best way to satisfy a faceted topic retrieval need is to retrieve the smallest set of documents that contains all of the facets, and thus that *S-rec* is the most natural measure to evaluate a faceted topic retrieval system.

Finding the minimum rank is an instance of the Minimum Set Cover problem and is therefore NP-Hard. To see this, consider the universe $\mathcal{F}$ of facets for a query Q. Define a document D_i as a subset of $\mathcal{F}$, i.e. a document contains a subset of facets. The minimum rank is equivalent to the size of the smallest subset of documents such that their union contains all elements of $\mathcal{F}$—exactly the Minimum Set Cover problem. However, the way facets are empirically distributed in documents allows for some heuristics. First, any document that contains a subset of the facets contained in another document can be eliminated from consideration. Second, if any facets always occur in documents separately from other facets, one of those documents must be part of the set. We can then find an upper bound on the minimum rank by taking documents in a greedy fashion according to the number of unsatisfied facets they satisfy. The size of the resulting set of documents is the minimum rank. Comparing this algorithm to exhaustive search suggests that it produces a very tight approximation, with an error of less than 0.5 on any individual set of facets.

One might think that focusing on the smallest set of documents would result in very long documents being preferred, perhaps because those are more likely to contain more facets. But because facets are so closely related to each other (by the definition of the task), we do not believe that this will happen; in fact, it is likely that more facets would be retrieved in short, very focused documents than in longer ones.

3.2.2 S-Precision

S-recall measures the ability of the system to find facets, but we would also like to know whether it ranks them well. Zhai et al. define S-precision at rank k as the minimum rank required by a perfect system to achieve recall of at least $rec@k$, divided by k: k'/k, where k' is the minimum rank at which *S-rec*@k could possibly be achieved. This can be understood by analogy to traditional precision by thinking of the number of relevant documents retrieved (the numerator of precision) as the minimum rank required to reach the recall at the same rank. This variant of precision has the same properties as traditional precision: it ranges from 0 to 1; it is greater when more unique facets have been retrieved; it approaches zero as $k \to \infty$.

Like finding the minimum optimal rank, calculating S-precision is NP-Hard. Again, the greedy approximation algorithm works very well in practice.

3.2.3 Redundancy

When a facet F_j occurs in the document at rank 2 after having already occurred in the document at rank 1, its appearance in document 2 is *redundant*. There is often a tradeoff between eliminating redundancy and retrieving the smallest set of documents that contain all the facets: less redundancy may require more documents to cover all the facets. Therefore we evaluate redundancy at rank k separately from recall and precision (Zhang et al. also argued for evaluating redundancy separately [17]). Redundancy is the average number of times each facet is duplicated up to rank k (if there are no relevant documents ranked above k, redundancy is undefined). Between two systems with the same *S-rec*, the one with lower redundancy should generally be preferred, but lower redundancy is not by itself a reason to prefer a system.

In Figure 1, four of the 14 facets would be retrieved more than once in the smallest possible set. It is impossible to cover all 14 facets without redundancy.

4. FACETED TOPIC RETRIEVAL MODELS

The *Probability Ranking Principle* is a well-known guideline for ranking documents in standard IR tasks such as ad hoc retrieval [13]. It says that optimal performance is achieved when documents are ranked in decreasing order of probability of relevance. It therefore provides guidance for building retrieval systems: systems that do a better job at predicting relevance will perform better by precision and recall measures.

The PRP assumes that documents are independently relevant [9]. This is not the case in faceted topic retrieval, as Figure 1 suggests. If the top two documents are LAT20040204-.0043 and APE20040401.0108, there is no additional benefit to retrieving LAT20040430.0068 at rank 3, even though it is relevant to the topic. S-recall captures this by rewarding a system for retrieving a different, non-redundant, relevant document at rank 3, while traditional recall does not.

In this section we describe two well-known models for novelty ranking, and propose two new models. The two well-known models are a heuristic approach and a probabilistic analogue; the new models follow the same pattern.

4.1 Maximal Marginal Relevance

A natural approach to this problem is *maximal marginal relevance*, defined by Goldstein & Carbonell [4]. As the name suggests, MMR is a greedy ranking method that chooses the ith document in a ranking according to a combination of its similarity to the query and its similarity to the documents ranked at positions 1 to $i-1$:

$$MMR(D_i, Q) = \alpha sim_1(D_i, Q) - (1-\alpha) \max_{1 \le j < i} sim_2(D_i, D_j)$$

where sim_1 is a standard query-document scoring function, sim_2 is a similarity function between documents, and α is a parameter. When $\alpha = 1$, a ranking by MMR is equivalent to a ranking by the query-document similarity. MMR is a simple but effective approach to novelty ranking, and therefore an obvious approach to faceted topic retrieval.

4.2 Probabilistic Interpretation of MMR

Zhai et al. proposed a probabilistic interpretation of MMR. Documents are scored on the basis of two probabilities: a probability of relevance $P(rel|D_i)$ and a probability of containing novel information $P(new|D_i)$. These two probabilities are combined together in a scoring function as:

$$\begin{aligned} s(D_i|D_1, ..., D_{i-1}) = {} & c_1 P(rel|D_i) P(new|D_i) \\ & + c_2 P(rel|D_i) P(\overline{new}|D_i) \\ & + c_3 P(\overline{rel}|D_i) P(new|D_i) \\ & + c_4 P(\overline{rel}|D_i) P(\overline{new}|D_i). \end{aligned}$$

Zhai et al. argue that there is no cost to presenting a novel relevant document ($c_1 = 0$) and that the cost of presenting a nonrelevant document is unaffected by whether that document is novel or not ($c_3 = c_4$), resulting in the final rank-equivalent scoring function:

$$s(D_i|D_1, ..., D_{i-1}) = P(rel|D_i)\left(1 - \frac{c_3}{c_2} - P(new|D_i)\right).$$

The ratio c_3/c_2 can be replaced with a single parameter ρ. The problem thus reduces to estimating $P(rel|D_i)$ and $P(new|D_i)$. $P(rel|D_i)$ is naturally estimated using a language model. Zhai et al. present several methods for estimating $P(new|D_i)$, the most effective of which is called *AvgMix*. The AvgMix estimate is calculated by maximizing the log-likelihood of observing D_i after sampling n words from a mixture of an "old" model (i.e. of a previously-ranked document) and a background model with respect to mixing parameter λ. Greater λ means D_i is less likely to model the previously-ranked documents, and therefore more likely to be novel. This mixing parameter is found for each document at ranks 1 through $i-1$, then averaged for a final estimate of $P(new|D_i)$.

4.3 Greedy Result Set Pruning

Instead of greedily ranking documents using an estimate of novelty, we could rank documents by their similarity to the query, then prune that ranking of the documents that are most similar to other documents in the ranking. In this method, we simply step down the ranked list of documents (in order of relevance) and prune documents with similarity greater than some threshold θ. I.e., at rank i, we remove any document D_j, $j > i$, with $sim_2(D_j, D_i) > \theta$. This approach may result in different rankings than MMR, since it uses query similarity and novelty in two separate steps rather than combining them in one.

4.4 A Probabilistic Set-Based Approach

Our set-based formulation of S-recall suggests a set-based ranking principle for faceted topic retrieval: retrieve the *set* of documents that maximizes the likelihood of capturing all of the facets. This can be visualized by generalizing Figure 1 to a graph in which instead of 0-1 edges between documents and facets, each edge has a weight representing the probability that each document contains every possible facet in the universe. The goal of the faceted topic retrieval system is to find the smallest set of documents that "covers" the facet space with highest probability. Figure 2 shows the probabilistic facet graph. In this example, instead of 14 known facets there is a (countably) infinite universe of possible facets, of which the 17 shown have highest probability, and the "true" 14 are a subset of those. Every document has some probability of containing every facet; the thickness of the edge reflects the strength of the belief.

Suppose we have a particular hypothetical set of facets F and a set of documents D. Denote the probability that D contains F as $P(F \in D)$. This is the probability we wish to estimate, and ultimately maximize over sets D and F. As the equation for S-recall suggests, this is a probabilistic OR problem: F_j can be in document D_1, or it can be in document D_2, or in document D_3, and so on. We do not require that it be in all of them, only that it be in at least one. The well-known "sum rule" of probabilities tells us that

$$\begin{aligned} & P(F_j \in \{D_1, D_2\} = P(F_j \in D_1 \cup F_j \in D_2) \\ & \quad = P(F_j \in D_1) + P(F_j \in D_2) - P(F_j \in D_1, F_j \in D_2). \end{aligned}$$

As the number of documents grows, the number of clauses in the OR statement grows, and the number of terms in the expanded probability grows exponentially. With even a small set of documents it is infeasible to calculate. However, if we assume that a facet occurs in documents independently (i.e. $P(F_j \in D_1, F_j \in D_2) = P(F_j \in D_1)P(F_j \in D_2)$), we can estimate the probability as follows:

$$P(F_j \in \{D_1, D_2\}) = 1 - (1 - P(F_j \in D_1))(1 - P(F_j \in D_2)).$$

In general, then, the probability that a facet F_j occurs in at

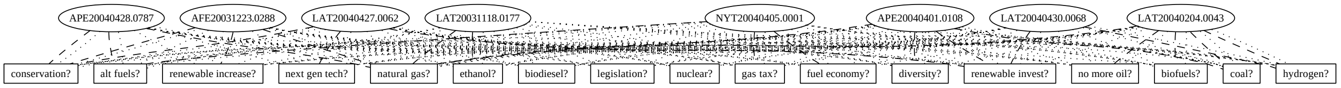

Figure 2: Example probabilistic document-facet graph. Every document contains every facet with some probability; the darker edges indicate a higher probability. Note that there are additional facets that were not in Figure 1; these represent other possible facets that do not actually appear in these documents.

least one document in a set D is

$$P(F_j \in D) = 1 - \prod_{i}^{n}(1 - P(F_j \in D_i))$$

and the probability that all of the facets in a set F are captured by the documents D is

$$P(F \in D) = \prod_{j=1}^{m} P(F_j \in D) = \prod_{j=1}^{m} 1 - \prod_{i}^{n}(1 - P(F_j \in D_i)).$$

Note that this involves a second independence assumption: that facets occur in documents independently of one another. Because the retrieval is set-based, these independence assumptions will not controvert the statement of the problem that facets are correlated across documents.

It is fairly easy to see that maximizing $P(F \in D)$ with a subset D of corpus $\mathcal{C}$, $|D| = k$, directly results in maximizing *S-rec*@k, much as maximizing $P(R|Q, D)$ for relevant documents directly results in maximizing precision and recall. S-precision will not necessarily be maximized, nor redundancy minimized, but there is a tradeoff involved: because the set of documents that maximize recall at rank 2 is not necessarily contained in the set of documents that maximize recall at rank 3.[1], we must make some decision to try to optimize recall at a particular rank or to optimize the number of new facets retrieved at each rank. This decision comes into play at the optimization phase.

A faceted topic retrieval engine using this model must do three things. First, it must hypothesize a set of facets. Second, for each facet F_j, it must estimate the probability $P(F_j \in D_i)$ that it occurs in each document D_i. Third, it must have a way to select the smallest set of documents that is most likely to contain all the facets. We will consider each of these in turn.

4.4.1 Hypothesizing Facets

Given only a short query, the system must be able to produce some hypothetical set of facets against which to score documents. There are various ways we could do this, including clustering, topic modeling, relevance modeling, phrase extraction, and so on. Some approaches may be supervised, others unsupervised. We have chosen to evaluate two unsupervised probabilistic methods: topic modeling with LDA and relevance modeling. Others will be left for future work.

In both cases we will assume that we have been given a set of documents from which to "extract" some facets. Instead of trying to extract any particular word or phrase to use as a facet, we will instead build a "facet model" $P(w|F)$. Our hope is that two different facet models will capture something about the vocabulary associated with different facets by assigning higher probabilities to different terms. In our example above, we may have a facet model that corresponds to "biofuels" by giving higher probabilities to words that co-occur with "biofuel" more often than they co-occur with other terms, and one that corresponds to "gas tax" by giving higher probabilities to words that co-occur with "gas tax" more often than other terms. Then $P(D_i|F_{biofuel}) > P(D_i|F_{gas\ tax})$ suggests that document D_i is more likely to contain the "biofuels" facet than the "gas tax" facet (where $P(D_i|F_j) = \prod_{w \in D_i} P(w|F_j)$, i.e. it is a unigram language model).

A relevance model is a distribution of words $P(w|R)$ estimated from a set of relevant or retrieved documents [11]. Similarly, we will estimate m "facet models" $P(w|F_j)$ from a set of retrieved documents using the so-called RM2 approach described by Lavrenko and Croft [11]:

$$P(w|F_j) \propto P(w) \prod_{f_k \in F_j} \sum_{D_i \in D_{F_j}} P(f_k|D_i)P(w|D_i)p(D_i)/P(w)$$

where D_{F_j} is the set of documents relevant to facet F_j, f_k are the facet terms, $P(w) = \sum_{D_i \in D_{F_j}} P(w|D_i)P(D_i)$, and $P(w|D_i)$ is a smoothed estimate. Since we do not know the facet terms or the set of documents relevant to the facet, we will estimate them from the retrieved documents. We obtain m models from the top m retrieved documents by taking each document along with its k nearest neighbors as the basis for a facet model.

In the LDA approach, the "facets" are actually latent variables that are priors for document term occurrences [3]. Probabilities $P(w|F_j)$ and $P(F_j)$ are found through expectation maximization. Then we can find $P(D_i|F_j) = \prod P(w|F_j)$ and $P(F_j|D_i) \propto P(D_i|F_j)P(F_j)$, generally assuming a uniform prior on documents. Again, we limit the calculation to documents retrieved for a particular query; this may limit the ability of LDA to identify topics that represent facets.

Facet models can be built from the set of documents in an initial ad hoc-style retrieval, and thus both of these can be seen as query expansion/relevance feedback methods. But instead of a single expanded query, there are m, where m is the hypothesized number of facets. In this work we assume constant, manually-selected m from query to query; a full optimization would be over the number of facets as well as hypothesized facets and subsets of documents.

4.4.2 Estimating Document-Facet Probabilities

Both the facet relevance model and LDA model produce generation probabilities $P(D_i|F_j)$, i.e. the probability that sampling terms from the facet model F_j will produce document D_i. This is not a probability that a document *contains* a facet, which is what our model requires. However, much as the so-called query-likelihood $P(Q|D)$ is not a probability of relevance yet is useful for ranking documents, these

[1] If this type of containment were necessarily true, the evaluation problems would not be NP-Hard; they would in fact be solvable in polynomial time by the greedy algorithm. The fact that the greedy algorithm only gives an approximation provides evidence for this claim.

probabilities may still be useful in a faceted topic retrieval system. We consider this an empirical question.

Since the probabilities are likely going to be very small, to avoid numerical errors we will rescale them to a range more suited to the binomial containment variable. We elected to linearly scale them to the range $[0.25, 0.75]$. In a practical sense, this defines the containment probability using generation probability as a feature. We could easily incorporate additional features in a supervised training phase; this is left for future work.

4.4.3 Maximizing Likelihood

In Section 4 we presented the probability that a set of facets occurs in a set of documents. Above we discussed ways to choose facets; we also need a way to select a subset of documents. Let $y_i = 1$ if document D_i is selected and 0 otherwise. Then we can define the likelihood function:

$$L(y|F, D) = \prod_{j=1}^{m} 1 - \prod_{i=1}^{n} (1 - P(F_j \in D_i))^{y_i}. \qquad (1)$$

If $y_i = 0$ then $(1 - P(F_j \in D_i))^{y_i} = 1$ and D_i therefore has no effect on the likelihood. This expression is maximized with the constraint that $\sum y_i \leq k$, i.e. the total number of documents taken is no more than a hypothesized minimum number required to cover the facets.

Note that maximizing $L(y)$ is a 0-1 integer programming problem, which is NP-Hard in general. We can approximate the solution in various ways. Perhaps the most intuitive is analogous to our greedy algorithm for S-recall: greedily take the document that maximizes the likelihood conditional on the documents that have already been taken. This ensures that the first document taken produces the greatest expected number of facets, the second produces the greatest expected number of facets that are different from those provided by the first, and so on. It also provides a natural ranking of documents in order of their selection, and does not require any estimate of k. In execution it is very similar to MMR.

The greedy approach, while accounting for both diversity and redundancy, cannot necessarily maximize diversity. We therefore propose a simpler set-based approximation scheme: for each facet F_j, take the document that maximizes its probability $\arg\max_i P(F_j \in D_i)$. Note that this provides no ranking of the documents selected, so we rank them by their original ad hoc retrieval scores.

An alternative approach is to relax y to a vector of real numbers rather than 0-1 integers. This results in a likelihood function that is convex and differentiable, and thus can be solved with conjugate gradient descent methods. The constraint $\sum y_i \leq k$ is no longer valid in this approach, since y_i is no longer an indicator for the presence or absence of a document. Without that constraint, Eq. 1 is actually maximized by giving maximum score to every document; we introduce a penalization term $\lambda \sum y_i^2 = \lambda||y||$ to ensure that maximum scores are assigned to those documents most likely to contain facets:

$$\log L(y) = \sum_{j=1}^{m} \log \left(1 - \exp \sum_{i=1}^{n} y_i \log(1 - P(F_j \in D_i)) \right) + \lambda||y||.$$

After maximization, the scores y_i provide a natural ranking of documents.

5. EXPERIMENT

In this section we describe experimental faceted topic retrieval systems and apply them to data annotated with facets.

5.1 Data

There is no standard corpus for faceted topic retrieval. Allan et al. investigated the relationship between system performance and human performance on a faceted topic retrieval task [2]; we obtained the queries and facet judgments used in that work. The data consists of 61 topics, each with a short (3-6 word) query, and judgments of relevance to documents in a subset of the TDT5 corpus. A few of the queries are ambiguous and duplicated with different statements of information need (*a la* Sanderson [14]). For example, the query "Bush visits" appears twice, once in the context of foreign leaders that traveled to the U.S. to visit George W. Bush, and once in the context of places that Bush visited during his time as president.

There are three levels of judgment: a binary relevance judgment for the document; for each relevant document, a list of facets it contains; and for each facet, a passage in the document that supports its relevance to that facet. The documents judged are the top 130 retrieved by a query-likelihood language model for the short query. Since few documents were judged, it is very possible that facets exist in the corpus but do not appear in the judged documents. To ensure we have judgments on all ranked documents, we will only rerank these 130 documents for each query.

Sixty topics were annotated by two assessors (one was annotated by only one assessor; this was discarded). The statement of the information need contained guidelines on how to assess facets, but within those guidelines assessors were free to name the facets however they liked. On average, there were 44.7 relevant documents per query; each of those contained 4.3 facets. There were 39.2 unique facets on average, for an average of just under one unique facet per relevant document. Agreement about relevance was quite high (72% of all relevant documents were judged relevant by both assessors), but there was substantial disagreement about the number of facets per query (a difference of 8 facets on average). Assessor agreed about the number of facets per relevant document within one facet.

5.2 Retrieval Engines

We implemented all the models described in Section 4 using the Lemur toolkit, as well as standard language modeling and language modeling plus pseudo-feedback with relevance models. Whenever possible, we have used the same similarity or scoring functions between models to ensure the fairest possible comparison. Specifically, our models are:

- LM baseline: a basic query-likelihood (Dirichlet smoothing; $\mu = 1000$) run with no facet model.
- RM baseline: a pseudo-feedback run with relevance modeling and no facet model.
- MMR: maximal marginal relevance with query similarity scores from the LM baseline and cosine similarity for novelty. Query-likelihood scores are re-scaled to $[0, 1]$ to make them compatible with cosine similarities.
- AvgMix: the probabilistic MMR model using query-likelihood scores from the LM baseline and the AvgMix novelty score.

- Pruning: removing documents from the LM baseline based on cosine similarity to lower-ranked documents.

- FM: the set-based facet model described in Section 4.4.1.

For the set-based model, we have two different ways to hypothesize facets and score documents.

- FM-RM refers to the facet relevance model described above. Each of the top m documents and their K nearest neighbors becomes a "facet model" $P(w|F_j)$—a truncated (v-term) relevance model constructed from the documents. Then we compute the probability $P(D_i|F_j)$ for each document and facet model; these are converted to a probability $P(F_j \in D_i)$ by linear transformation to the range $[0.25, 0.75]$.

- FM-LDA uses subtopics discovered using LDA. This provides $p(z_j|D_i)$ for each document D_i and each "subtopic" z_j; these were used as the facet-document scores. We extracted 50 subtopics.

Finally, we performed a manual "oracle" experiment using one assessor's facet labels as queries to score documents against using query-likelihood. Like the FM-RM, these scores were rescaled to $[0.25, 0.75]$ and the optimization methods in Section 4.4.3 applied. This provides a loose upper bound on the performance of FM.[2]

5.3 Optimization in the Set-Based Model

As discussed above, the likelihood maximization is a 0-1 integer programming problem. This is NP-Hard in general. Instead of trying to solve it directly, we tested several different approximate solutions:

- For each facet F_j, take the document D_i with maximum $P(F_j \in D_i)$. We call this *max-set* because it produces a set of documents. We rank the documents by their original query-likelihood score.

- Greedily take the document D_i that maximizes the likelihood conditional on documents taken in previous iterations, i.e. $\arg\max_i L(y_i|y_1, y_2, ..., y_{i-1}, F, D)$. We call this the *marginal likelihood* method.

- Relax y to a real-valued vector and solve using conjugate gradient descent. We call this the *relaxed optimization* method.

Our results below use max-set; we compare that to the other approaches in Section 6.1.

5.4 Experiment and Evaluation

We used five-fold cross-validation to train and test systems, and to obtain results for all 60 queries for each model. We divided the 60 queries into five folds of 12 queries each. The 48 queries in four folds are used as a training set to select model parameters such as $\alpha, \rho, \theta, (m, K, v)$ (for MMR, AvgMix, pruning, and set models, respectively). These parameters are used to obtain ranked results on the remaining 12 queries. The query splits were chosen randomly in advance so that all experiments used the same training and testing data.

For each method we report S-recall at the minimum optimal rank *S-rec* (which ranges from 0 to 1, larger values indicating better performance), redundancy at the minimum optimal rank (which has a minimum of zero but no upper bound; smaller is better), and mean average precision (MAP) using the document-level relevance judgments. We also show 11-point interpolated S-precision/S-recall curves.

The average minimum optimal rank is 10, which is a good match to the first page of results in web search engines. There is substantial variance over queries, however, with some having a minimum rank of 56 and others having a minimum rank of 1 (there is a single document that contains all the facets). Comparing our greedy algorithm to exhaustive search on a subset of topics shows that the greedy algorithm's approximation is not perfect but very close.

Since there were two assessors for each topic, we test hypotheses about differences between systems using a two-way within-subjects ANOVA on S-recall. A two-way ANOVA calculates the variance in a measurement of recall due to differences between systems and due to differences between assessors, as well as interactions between the two. We would like to see that the variance due to systems is significant and outweighs any other source of variance. If this is the case, the comparison is robust to differences in assessors. Ideally we would like to see that variance due to assessors is *not* significant, and in particular that the interaction is negligible.

6. RESULTS AND ANALYSIS

Table 1 shows S-recall, redundancy ratio, and mean average precision (MAP) for all systems described in Section 5.2. Among the seven automatic methods, result-set pruning gives the best overall results, though we note that there is no significant difference in the S-recalls of MMR, pruning, and FM-RM. All three retrieved about 44% of the facets, compared to roughly 40% by the two baselines and AvgMix, and only 15% by FM-LDA. All of the models (except FM-LDA) exhibited a fairly high degree of redundancy, duplicating each facet at least 0.5 times (on average) in the relevant documents retrieved by the minimum rank. MMR had the lowest redundancy, significantly lower than pruning and FM-RM. The very low redundancy of FM-LDA may be explained by the fact that it retrieved very few relevant documents; the recall-redundancy in Figure 3 gives a better sense of how redundancy varies with S-recall for each run.

The three best runs are significantly better than any of the others. Table 2 shows a summary of the results of an ANOVA to determine whether differences among the top five automatic runs are affected by assessor disagreement. Indeed, system differences are significant, while assessor differences are not, and there is negligible interaction between system and assessor (systems are not fitting to certain assessors).

The "manual" results in Table 1 provide some loose upper bounds. It retrieved 68% of the facets, but still retrieved each of them almost as many times as the facet model. This suggests that a fairly high degree of redundancy is inevitable. Many of the harder-to-find facets are only present in documents that contain easy-to-find facets; it is simply not possible to retrieve all of these without some redundancy. This run therefore suggests that lower redundancy is only super-

[2] It is a very loose bound because, as stated above, assessors could name facets however they liked. Any unusual names would cause poor scores. For example, one assessor used abbreviated labels such as "NDakota", "WashDC", "SCarolina"; no document could score well against such queries.

system	*S-rec*	redundancy	MAP
LM baseline	0.405	0.856	0.583
RM baseline	0.376	1.176	**0.617***
MMR	0.440	0.538	0.534
prob MMR	0.398	0.720	0.570
pruning	**0.444**	0.567	0.501
FM-RM	0.440	0.674	0.574
FM-LDA	0.153	**0.224***	0.285
manual	0.677	0.672	0.698

Table 1: S-recall and redundancy at the minimum optimal rank and average increase in S-recall from rank 1 to the minimum optimal rank for four faceted topic retrieval systems. Numbers are averaged over 60 topics with two sets of assessments each. The best automatic result for each column is in bold. An asterisk indicates statistical significance.

factor	*df*	*F*	*p*-value
system	4	5.615	0.000
assessor	1	1.018	0.317
system:assessor	4	1.689	0.153

Table 2: Two-way ANOVA results on S-recall for the LM baseline, MMR, AvgMix, pruning, and FM-RM. Differences between systems are significant while differences between assessors do not significantly affect the results. There is insignificant interaction between assessor and system.

ficially desirable; optimizing for redundancy may result in "harder" facets being missed. We explore this in Section 6.2 below.

Because it uses the set-based framework we presented in Section 4.4.1, this manual run also suggests that the set-based model is easily improved simply by improving the facet models: if a user provides some information about the facets, we can easily incorporate it into the facet models. This stands in contrast to MMR or AvgMix, which cannot incorporate such information as easily.

Figure 3 shows the 11-point S-precision/recall curves for seven systems. Five of them coincide closely, though the MMR, pruning, and FM-RM curves are clearly above the LM baseline and AvgMix curves. The FM-RM curve is above the others at both the highest and lowest recall levels, but not in between. The FM-LDA system significantly underperforms compared to the others. Figure 3 also shows redundancy increasing with S-recall. The LM baseline has the highest redundancy, followed closely by AvgMix. The FM is "middle-of-the-road", almost exactly in between all the other automatic runs. FM-LDA and pruning coincide closely over all S-recall values. MMR and the manual run coincide closely up to S-recall 0.5; after that the redundancy of MMR increases to match or exceed that of pruning.

We also computed basic mean average precision (MAP) using the document-level relevance judgments. Note that MAPs in Table 1 are high because every system was able to rank all judged documents. They should be considered upper bounds, though their relative ordering would not change with more judgments. All of the non-baselines had lower MAPs than the baselines, demonstrating the inadequacy of MAP for this task.

		optimization method		
measure	model	max-set	marginal	relax
S-rec	FM	**0.440***	0.392	0.388
	LDA	**0.153**	0.113	0.128
	manual	0.677	**0.723***	0.677
redundancy	FM	**0.674***	0.914	1.293
	LDA	0.224	**0.172***	1.142
	manual	**0.672***	0.978	1.254
MAP	FM	0.574	0.509	**0.625***
	LDA	0.285	0.277	**0.327***
	manual	0.698	0.702	**0.766***

Table 3: Results with different optimization methods. Bolded numbers are the best across the row. An asterisk indicates statistical significance by a within-subjects 2-way ANOVA ($p < 0.05$; assessor effects are not significant). The max-set approximation tends to provide the best diversity results, while the relaxation approach provides the best MAP.

6.1 Optimization

Table 3 compares different approximate optimization approaches for FM. The baselines do not require any optimization, so they are not shown. The max-set method gives significantly higher S-recall and lower redundancy for the facet model. For the LDA model, it gives better (but not significantly so) S-recall but also greater redundancy.

The marginal-likelihood approach for FM-RM gives similar results to AvgMix. Both may be seen as probabilistic variations of MMR, so this is not surprising. It is notable that marginal-likelihood provides the best S-recall for the manual facet model. Though the redundancy is higher than the max-set approach, this is probably a consequence of retrieving more relevant material. These results suggest that marginal-likelihood is successful when there is a high degree of confidence in the facets, but less so if not.

The improvement in MAP under the relaxed optimization approach was surprising to us. The relaxed approach tends to give similar weights to documents that are similar in the facet space; it is more optimal in this approach to give three identical documents each weights of 1/3 than to give one of them a weight of 1 and the other two 0. Thus we hypothesize that it is identifying low-ranked relevant documents that are "similar" in the facet space to higher-ranked documents and moving them up in the ranking. The fact that its redundancy tends to be higher supports this.

6.2 Additional Analysis

Because all of the subproblems of faceted topic retrieval are difficult, there are multiple points of failure: the hypothesized facets could be unrelated to actual facets; even if the hypothesized facets are good, the probability estimates $P(F_j \in D_i)$ could be bad; even if the probability estimates are good, the optimization techniques could be bad because of independence assumptions or because of bad approximations. In this section we will consider some of these questions.

Are the hypothesized facets anything like the actual facets? We looked at term probabilities for FM and LDA facet models to try to determine whether there was any relationship between hypothesized facets and actual facets. Though we

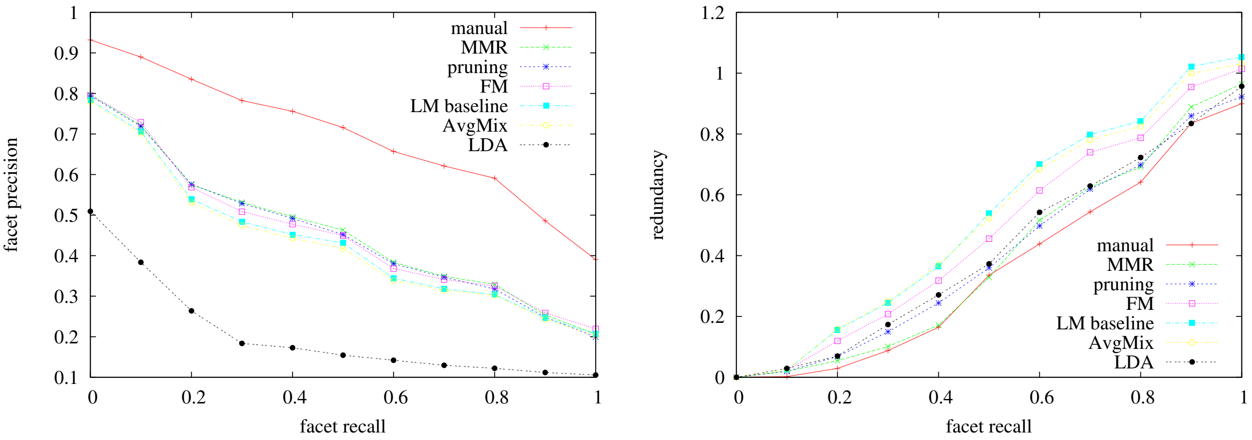

Figure 3: S-recall vs. S-precision (left) and redundancy (right) for seven models.

can only make subjective observations, the FM facet models do seem to capture actual facets in some cases. For example, for query 1 *oil-producing nations*, nearly all of the facet models could clearly be labeled with a nation or region; the first five correspond to "sub-Saharan African nations", "Chad", "Iraq", "Indonesia", and "Burma" based on the term distributions. But there are also overlapping models and duplicates, and some facets that do have representative models occur after enough duplicates that they are not ranked highly enough to be counted in *S-rec*.

On the other hand, for query 52 *disarm landmines land mines*, the facet models seem to correspond more to activist protests and demonstrations related to landmines rather than to strategies taken to disarm them (which is the stated user information need).

The LDA model performed poorly on average, possibly because of a lack of data to estimate topics or too much correlation between facets. There were a few topics for which it outperformed the baselines. Query 15 *Bush visits* is looking for foreign leaders that visited Bush in the U.S. Interestingly, looking at the term distributions suggests that the facet models are strongly associated with countries that visiting foreign leaders call home. For example, one of the topics gave high probability to terms such as *london, britain, blair*, etc. Query 20 *Bush visits* is instead looking for countries Bush visited; while LDA outperformed the baseline for both, it was by a much smaller degree for 20 than for 15.

For other queries such as query 51 *allies Israel*, it was difficult to identify how the topic models were differentiated. It seems likely that this is a result of having so few documents from which to estimate topic models.

How similar are the models in terms of facets they retrieve? We looked at the amount of overlap in facets retrieved by MMR, FM-RM, and pruning. Of all unique facets retrieved in the top 10 by MMR and FM-RM, 77% (476 of 613) were retrieved by both systems. For MMR and pruning, 82% (498 of 606) were retrieved by both. FM-RM and pruning agreed on 80% (491 of 610). The systems do display some differences.

Does any model do a better job on the "hard" facets? To answer this, we looked at pairs of models and identified the facets that were found by one model but not by the other. We then calculate the average number of documents these facets occurred in. For example, if MMR found the facet *United Arab Emirates* for the query *oil producing countries*, but FM-RM did not, we would look at the judgments and see that UAE appears in nine different documents. If FM-RM found *Kuwait* and MMR did not, we would look at the judgments and see that Kuwait appears in eight different documents. We would conclude that FM-RM did a very slightly better job at finding slightly harder facets for this query.

Over all 60 queries, the average number of appearances of facets retrieved in the top 10 documents by FM-RM but not by MMR is 2.77, whereas the average number of appearances of facets retrieved in the top 10 documents by MMR but not FM-RM is 4.13 (there were 613 unique facets found by both systems, with 137 found by one but not the other). FM-RM therefore seems to do a better job of finding "harder" facets. On the other hand, MMR seems to do better than pruning (2.54 average appearances for MMR versus 3.53 for pruning). These relationships hold consistently as the number of retrieved documents increases, suggesting that FM-RM is better able to find harder facets than MMR, which in turn is better than pruning.

Does true optimization of FM give better results? Though the optimization problem is NP-Hard, for small n it is feasible to try all $\binom{130}{n}$ document subsets to determine which is optimal. With $n = 2$ there are $\binom{130}{2} = 8385$ possible subsets for each query. The problem quickly becomes infeasible; there are two orders of magnitude more possibilities when $n = 3$.

We took the size-2 subset with the greatest log-likelihood and calculated S-recall at rank 2. We compared that to S-recall at rank 2 for our other optimization methods with each model. The result is that the true optimal set produces better results than any approximate set: a 16% improvement in the case of FM down to a 1% improvement in the manual model. This suggests that there is value in exploring other optimization methods.

Can independence assumptions in FM be relaxed? Does it make a difference? In Section 4 we made an independence assumption to keep the computation tractable, namely that facets occur in documents independently:

$$P(F_i \in \{D_1, D_2\}) = P(F_i \in D_1)P(F_i \in D_2).$$

We can relax this assumption slightly by using covariance to model dependence between two documents:

$$P(F_i \in \{D_1, D_2\}) = P(F_i \in D_1)P(F_i \in D_2) + Cov(D_1, D_2)$$

where $Cov(D_1, D_2)$ is calculated by summing over all facets, i.e. it is a measure of the similarity between the two documents in the facet space.

We followed the same procedure as above, calculating the likelihood over all subsets of size 2 and taking the set that produced the maximum. We compared to the size-2 set above. Overall there is not a great deal of difference in S-recall: a 6% decrease for FM, a 4% increase for LDA, and a 3% increase for the manual run. There is a 10% decrease in redundancy for FM, however (and negligible changes in the other two). Modeling dependence does have some effect with "true" optimization, then, though there is interaction with the hypothesized facets that is difficult to quantify.

7. CONCLUSION AND FUTURE WORK

We have defined a type of novelty retrieval task called *faceted topic retrieval*: retrieve the facets of an information need in a small set of documents to be presented to the user. We presented two novel models for it: one that prunes a standard retrieval ranking and one a formally-motivated probabilistic model. We demonstrated that both models are competitive with MMR, and outperform another probabilistic model—and all models outperform the traditional IR baselines. Additionally, an upper bound experiment suggests that our probabilistic model could easily and naturally incorporate information about facets provided by users or extracted from other sources to improve results, whereas the other models could not incorporate this information without some reformulation.

We have only scratched the surface of what is possible within this framework; there is ample opportunity for work on hypothesizing or averaging over facet sets of different sizes (using hierarchical clusters, for instance), estimating the most likely number of facets, using additional features of documents and relevant passages to estimate $P(F_j \in D_i)$ with supervised approaches, exploring other optimization functions, and so on. We believe our model will be applicable to other problems as well, including metasearch and multi-modal retrieval.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, H. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM '09*, pages 5–14.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proceedings of SIGIR*, pages 433–440, 2005.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

[4] J. Carbonell and J. Goldstein. The user of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336, 1998.

[5] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of SIGIR*, pages 429–436, 2006.

[6] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659–666, 2008.

[7] W. Dakka and P. G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. *Data Engineering, International Conference on*, 0:466–475, 2008.

[8] W. Goffman. On relevance as a measure. *Information Storage and Retrieval*, 2(3):201–203, 1964.

[9] M. D. Gordan and P. Lenk. A utility theoretic examination of the probability ranking principle in information retrieval. *JASIS*, 42:703–714, 1991.

[10] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR*, pages 23–28, 2000.

[11] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR*, pages 120–127, 2001.

[12] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of ICML '08*, pages 784–791.

[13] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304, Dec. 1977.

[14] M. Sanderson. Ambiguous queries: Test collections need more sense. In *Proceedings of SIGIR*, pages 499–506, 2008.

[15] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.

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

[17] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of SIGIR*, pages 81–88, 2002.