Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval

ChengXiang Zhai
Computer Science Department
University of Illinois at Urbana-Champaign

William W. Cohen
Center for Automated Learning and Discovery
Carnegie Mellon University

John Lafferty
School of Computer Science
Carnegie Mellon University

ABSTRACT

We present a non-traditional retrieval problem we call *subtopic retrieval*. The subtopic retrieval problem is concerned with finding documents that cover many different subtopics of a query topic. In such a problem, the utility of a document in a ranking is dependent on other documents in the ranking, violating the assumption of independent relevance which is assumed in most traditional retrieval methods. Subtopic retrieval poses challenges for evaluating performance, as well as for developing effective algorithms. We propose a framework for evaluating subtopic retrieval which generalizes the traditional precision and recall metrics by accounting for intrinsic topic difficulty as well as redundancy in documents. We propose and systematically evaluate several methods for performing subtopic retrieval using statistical language models and a maximal marginal relevance (MMR) ranking strategy. A mixture model combined with query likelihood relevance ranking is shown to modestly outperform a baseline relevance ranking on a data set used in the TREC interactive track.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models—*language models, dependent relevance*

General Terms

Measurement, Algorithms

Keywords

Subtopic retrieval,maximal marginal relevance, language models

1. INTRODUCTION

The notion of relevance is central to many theoretical and practical information retrieval models. Traditional retrieval models assume that the relevance of a document is *independent* of the relevance of other documents. This makes it possible to formulate the retrieval problem as computing the relevance (or some correlated metric) for each document separately, and then ranking documents by probability of relevance [11]. In reality, however, this independent relevance assumption rarely holds; the utility of retrieving one document, in general, may depend on which documents the user has already seen. As an extreme example, a relevant document may be useless to a user if the user has already seen another document with the same content. Another example is when the user's information need is best satisfied with several documents working together; in this case, the value of any single document may depend on what other documents are presented along with it. Some of the issues concerning ranking interdependent documents are discussed in [11, 13].

In this paper, we study the *subtopic retrieval* problem, which requires modeling *dependent* relevance. The subtopic retrieval problem has to do with finding documents that cover as many *different* subtopics of a general topic as possible. For example, a student doing a literature survey on "machine learning" may be most interested in finding documents that cover representative approaches to machine learning, and the relations between these approaches. In general, a topic often has a structure that involves many different subtopics. A user with a high-recall retrieval preference would presumably like to cover all the subtopics, and would thus prefer a ranking of documents such that the top documents cover different subtopics.

The same problem, called "aspect retrieval," is investigated in the interactive track of TREC, where the purpose is to study how an interactive retrieval system can best support a user gather information about the different aspects of a topic [9, 10, 5]. Here we study the task of automatically ranking documents so as to give good subtopic retrieval. In other words, we retain the basic "query in—ranked list out" model used in traditional retrieval, but seek to modify the ranking so as to include documents relevant to many subtopics.

Clearly, methods based on a traditional relevance-based ranking are unlikely to be optimal for such a problem. Moreover, traditional evaluation metrics are also inappropriate for this new retrieval task. We present an initial study of this new problem, describing evaluation metrics, possible methods, and experimental results with these methods.

2. DATA SET

In order to measure how well a ranking covers different subtopics of some high-level topic, we must have judgments that tell us which documents cover which subtopics. Fortunately, the TREC interactive track has accumulated many such judgments over the three years when the task was evaluated (TREC-6, TREC-7, and TREC-8). We collect all these judgments and use them for our analysis

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

and experiments. The document collection used in the interactive track is the Financial Times of London 1991-1994 collection (part of the TREC-7 ad hoc collection). This collection is about 500MB in size and contains 210,158 documents, with an average document length of roughly 400 words. Six to eight new topics were introduced each year, for a total of 20 topics, all of which were used in the work reported here. These interactive track topics were formed by slightly modifying the original ad hoc TREC topics, typically by removing the "Narrative" section and adding an "Instance" section to explain what a subtopic means for the topic. We generate the query for each topic by concatenating the title and the description of the topic.[1]

The following is an example query from TREC7 interactive track (number 392i).

```
Number:   392i
Title:    robotics
Description:
  What are the applications of robotics in the world
  today?

Instances:
  In the time alloted, please find as many DIFFERENT
  applications of the sort described above as you can.
  Please save at least one document for EACH such
  DIFFERENT application. If one document discusses
  several such applications, then you need not save
  other documents that repeat those, since your goal
  is to identify as many DIFFERENT applications of
  the sort described above as possible.
```

For each topic, the TREC (NIST) assessors would read a pool of documents submitted by TREC participants, and gradually identify a list of instances (i.e., subtopics) and record which documents contain or cover which instances. For example, for the sample topic 392i shown above, they identified 35 different subtopics, some of which are shown below:

```
  1  'clean room' applications in healthcare &
     precision engineering
  2  spot-welding robotics
  3  controlling inventory - storage devices
  ... ...
```

For this topic, the judgment for each document can be represented as a bit vector with 35 bits, each indicating whether the document covers the corresponding subtopic. In our data set, the number of subtopics (i.e., the range of vector lengths) ranges from 7 to 56, with an average of 20. The number of judged relevant documents available also differs for different topics, with a range of 5 to 100 and an average of about 40 documents per topic. There are also some judgments of non-relevant documents. We did not use these judgments; instead, we assume that any unjudged document is non-relevant, and therefore covers no relevant subtopic. This is a strong assumption, but our hope is that this biased evaluation will still be useful for comparing different rankings. More details about these data and the interactive track can be found in [9, 10, 5].

Note that the granularity of subtopics and the criteria to judge whether a document covers a subtopic are inevitably vague and subjective. A binary judgment also means that a document is assumed to either cover or not cover a subtopic, while in reality, the coverage may be somewhere in between.

3. EVALUATION METRICS

We wish to explore methods for producing a *ranked list* which performs well on the subtopic retrieval task. It is not immediately obvious how one should evaluate such a ranking. Intuitively, it is desirable to include documents from many different subtopics early in the ranking, and undesirable to include many documents that redundantly cover the same subtopics.

[1] While we have not explored it, a structured query (cf. [6]) can potentially be formulated to include keywords for different subtopics.

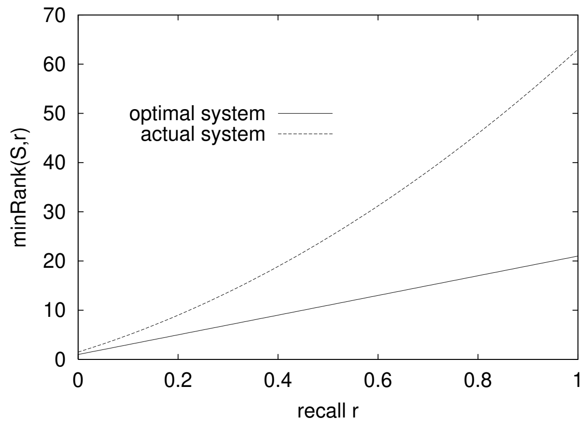

Figure 1: Typical curves for the functions $\mathrm{minRank}(\mathcal{S}, r)$ and $\mathrm{minRank}(\mathcal{S}_{opt}, r)$, defined as the minimal rank K at which subtopic recall of r is reached for system $\mathcal{S}$ and an optimal system $\mathcal{S}_{opt}$. Subtopic precision is defined as the ratio of $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ and $\mathrm{minRank}(\mathcal{S}, r)$.

One natural way to quantify success according to the first goal—of covering many different subtopics quickly—is to measure the number of different subtopics covered as a function of rank. More precisely, consider a topic T with n_A subtopics $A_1 \ldots, A_{n_A}$, and a ranking $d_1, \ldots, d_m$ of m documents. Let $\mathrm{subtopics}(d_i)$ be the set of subtopics to which d_i is relevant. We define the *subtopic recall (S-recall) at rank* K as the percentage of subtopics covered by one of the first K documents, i.e.,

$$\textit{S-recall at } K \equiv \frac{|\cup_{i=1}^{K} \mathrm{subtopics}(d_i)|}{n_A}$$

3.1 Accounting for intrinsic difficulty

Clearly it is desirable for subtopic recall to grow quickly as K increases. However, it is not at all clear what constitutes a "good" level of recall for a particular topic T. For example, consider two topics T_1 and T_2. For topic T_1, there are $M/2$ relevant documents and $M/2$ subtopics, and every document d_i covers exactly one distinct subtopic a_i. For topic T_2, there are $M/2$ relevant documents but M subtopics, and every document d_i covers subtopics $a_i, a_{M/2}, \ldots, a_M$. For both T_1 and T_2, the ranking $d_1, d_2, \ldots, d_{M/2}$ is clearly the best possible: however, subtopic recall for small rankings is much better for T_2 than for T_1. Similarly, for any natural measure for redundancy (the degree to which documents in a ranking repeat the same subtopics) the ranking for T_2 would appear much worse than the ranking for T_1.

This example suggests that for a measure to be meaningful across different topics, it must account for the "intrinsic difficulty" of ranking documents in a topic. We propose the following evaluation measure. If $\mathcal{S}$ is some IR system that produces rankings and r is a recall level, $0 \leq r \leq 1$, we define $\mathrm{minRank}(\mathcal{S}, r)$ as the minimal rank K at which the ranking produced by $\mathcal{S}$ has S-recall r. We define the *subtopic precision (S-precision) at recall* r as

$$\textit{S-precision at } r \equiv \frac{\mathrm{minRank}(\mathcal{S}_{opt}, r)}{\mathrm{minRank}(\mathcal{S}, r)}$$

where $\mathcal{S}_{opt}$ is a system that produces the *optimal* ranking that obtains recall r—i.e., $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ is the smallest K such that some ranking of size K has subtopic recall of r.

The idea of comparing performance to a theoretical optimal is not new [6]; however, this formulation of the comparison has some nice properties. Specifically, we claim that subtopic recall and precision, as defined above, are natural generalizations of ordinary recall and precision, in the following sense: if $\mathrm{minRank}(\mathcal{S}, r)$ were defined in terms of ordinary recall rather than subtopic recall, then ordinary precision could be defined as the ratio of $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ to $\mathrm{minRank}(\mathcal{S}, r)$.

To see this, consider the hypothetical curves for $\mathrm{minRank}(\mathcal{S}, r)$ and $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ shown in Figure 1. Suppose that $\mathcal{S}$ and $\mathcal{S}_{opt}$ are ordinary retrieval systems, and $\mathrm{minRank}$ is defined in terms of ordinary recall. Since $\mathcal{S}_{opt}$ orders all the relevant documents first, $\mathrm{minRank}(\mathcal{S}_{opt}, r) = r \cdot n_R$ (where n_R is the number of relevant documents for the topic). Now consider a non-optimal system $\mathcal{S}$ that has precision p and recall r in the first K_r documents. Since recall is r, $\mathcal{S}$ retrieves rn_R relevant documents in the first K_r, and its precision is $p = rn_R/K_r = \mathrm{minRank}(\mathcal{S}_{opt}, r)/\mathrm{minRank}(\mathcal{S}, r)$.

The hypothetical curves in Figure 1 are consistent with the performance of ordinary ranked retrieval systems: $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ grows linearly, and $\mathrm{minRank}(\mathcal{S}, r)$ becomes more gradually distant from the line for the optimal system, reflecting the fact that precision decreases as recall increases. Since the shape of $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ is predictable for ordinary retrieval, it is not necessary to explicitly account for it in measuring performance. For subtopic retrieval, however, $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ may have a more complex shape.

As concrete examples, the left-hand graphs in Figures 3 and 4 show subtopic recall and subtopic precision for various ranking schemes, interpolated over 11 points in the usual way, and averaged over all 20 topics in our test suite.

The S-precision and S-recall metrics are broadly similar to the cumulated gain (CG) measure proposed by Jarvelin and Kekalianen [6]. However, the CG measure assumes the gain of each document to be *independent* of other documents, and thus is insufficient for our purposes; in contrast, the "gain" of each document in the *S-precision* metric depends on other documents.

3.2 Penalizing redundancy

Intuitively, it is undesirable to include many documents that redundantly cover the same subtopics; however, this intuition is not accounted for in the measures of subtopic recall and precision.

One way to penalize redundancy is to include an explicit measure of the cost of a ranking. We let the *cost* of a ranking be defined as

$$\begin{aligned} \mathit{cost}(d_1, \ldots, d_K) &\equiv \sum_{i=1}^{K} (a|\mathrm{subtopics}(d_i)| + b) \\ &= a \sum_{i=1}^{K} |\mathrm{subtopics}(d_i)| + Kb \end{aligned}$$

Here b is the cost of presenting a document d_i to a user, and a is the incremental cost to the user of processing a single subtopic in d_i.

Proceeding by analogy to the measure introduced above, we define $\mathrm{minCost}(\mathcal{S}, r)$ to be the minimal cost C at which the ranking produced by $\mathcal{S}$ has S-recall r. We then define the *weighted subtopic precision (WS-precision) at recall level* r to be

$$\textit{WS-precision at } r \equiv \frac{\mathrm{minCost}(\mathcal{S}_{opt}, r)}{\mathrm{minCost}(\mathcal{S}, r)}$$

where again $\mathcal{S}_{opt}$ produces the optimal (lowest-cost) ranking that obtains recall r. Note that S-precision is a special case of WS-precision where $b = 1$ and $a = 0$. In this paper we will use costs of $a = b = 1$ for WS-precision.

Again, as concrete examples, the right-hand graphs in Figures 3 and 4 show subtopic recall and weighted subtopic precision for various ranking schemes.

Greedy Ranking Algorithm
Inputs: Set of unranked documents U; ranking size K
for $i = 1, 2, \ldots, K$ **do**
  $d_i = \arg\max_{d \in U} \mathit{value}(d; d_1, \ldots, d_{i-1})$
  $U = U - \{d_i\}$
endfor
return the ranking $\langle d_1, \ldots, d_K \rangle$

Figure 2: A generic greedy ranking algorithm

3.3 On computing the metrics

Computing S-precision and WS-precision require computing the optimal values $\mathrm{minRank}(\mathcal{S}_{opt}, r)$ or $\mathrm{minCost}(\mathcal{S}_{opt}, r)$. Unfortunately, this is non-trivial, even given relevance judgments. Indeed, it can be reduced to a minimum set-covering problem, which is NP-hard. Fortunately, the benchmark problems are of moderate size and complexity, and the minimum set cover can often be computed quite quickly using simple pruning heuristics. Furthermore, a simple greedy approximation seems to obtain results nearly indistinguishable from exact optimization, except at the highest recall values for $\mathrm{minCost}$.[2] In the evaluations of this paper, we used exact values of $\mathrm{minRank}$ for all queries. We used exact values of $\mathrm{minCost}$ for all queries but one (query 352i), and used a greedy approximation to $\mathrm{minCost}$ for query 352i.

4. SUBTOPIC RETRIEVAL METHODS

Since it is computationally complex to find an optimal ranking for the subtopic retrieval problem, even when the subtopics are known, some kind of approximation is necessary in practice. A natural approximation is a greedy algorithm, which ranks documents by placing at each rank i the document d_i that is "best" for that rank relative to the documents before it in the ranking. A generic version of this greedy algorithm is shown in Figure 2.

The key here is to appropriately define the *value* function–i.e., to quantify the notion of a "best" document d_i for rank i. Intuitively, d_i should cover many subtopics not covered by the previous documents $d_1, \ldots, d_{i-1}$, and few of the subtopics covered by the previous documents. Of course, one cannot compute such a metric explicitly in a value function, since the subtopics are not known to the retrieval system—only the initial query topic. Such an evaluation metric must therefore be based on a subtopic model.

An alternative to explicitly modeling subtopics is to use a similarity function that only implicitly accounts for subtopic redundancy. One such similarity-based approach is the maximal marginal relevance (MMR) ranking strategy [2]. MMR instantiates

[2]It is known that set cover is hard to approximate up to a logarithmic factor, and that the greedy algorithm achieves this factor [3]. For the 20 topics considered here, however, the greedy algorithm's performance actually is much better: for the 19 queries for which $\mathrm{minCost}$ could be computed exactly, the WS-precision of the greedy approximation is more than 99.6% for all recall values up to 0.9, and for recall 1.0, the WS-precision of the greedy approximation is 84%. Code implementing the exact and approximate greedy set covering algorithms is available on request from the authors.

the greedy algorithm of Figure 2 using the value function

$$\begin{aligned}&value_{MMR}(d; d_1, \ldots, d_{i-1}) = \\ &\quad \alpha Sim_1(d, Q) - (1-\alpha) \max_{j<i} Sim_2(d, d_j)\end{aligned}$$

where Q is the original query, α is a parameter controlling the relative importance of relevance and novelty, Sim_1 is a typical retrieval similarity function, and Sim_2 is a document similarity function that is intended to capture redundancy (or equivalently novelty).

Here we will study both novelty and relevancy in the language modeling framework. First, we will present two ways to measure the novelty of a document, one based on the KL-divergence measure, and another based on a simple mixture model. We will then discuss how to combine novelty and relevance in a cost function.

4.1 Novelty and Redundancy Measures

Let $\{\theta_1, ..., \theta_{i-1}\}$ be the unigram language models for $i-1$ previously selected documents, which we refer to as *reference language models*. Consider a candidate document d_i and the corresponding language model θ_i. Our goal is to define a novelty score $value_N$ for which $value_N(\theta_i; \theta_1, ..., \theta_{i-1})$ will indicate how much novel information document d_i contains.

4.1.1 Single Reference Topic Model

Let us first consider the simplest case, where we have a single reference model θ_O (where the O subscript indicates "old"). Suppose θ_N is the new document model. How do we define $value_N(\theta_N; \theta_O)$?

Notice that novelty is an *asymmetric* measure: we are interested in measuring the information in θ_N which is new relative to θ_O, not the other way around. For unigram language models, a natural asymmetric distance measure is the KL-divergence $D(\theta_N||\theta_O)$, which can be interpreted as the inefficiency (e.g., in compression) due to approximating the true distribution θ_N with θ_O. This leads to a value function of $value_{KL}(\theta_N; \theta_O) = D(\theta_N||\theta_O)$.

Another plausible novelty measure is based on a simple mixture model. Assume a two-component generative mixture model for the new document, in which one component is the old reference topic model and the other is a background language model (e.g., a general English model). Given the observed new document, we estimate the mixing weight for the background model (or the reference topic model), which can then serve as a measure of novelty or redundancy. The estimated weight can be interpreted as the extent to which the new document can be explained by the background model as opposed to the reference topic model. A similar idea, but with three-component mixture models, has been explored recently to measure redundancy in information filtering [16].

More formally, let θ_B be a background language model with a mixing weight of λ. The log-likelihood of a new document $d = w_1...w_n$ is

$$l(\lambda|d, \theta_O) = \sum_{i=1}^{n} \log((1-\lambda)p(w_i \mid \theta_O) + \lambda p(w_i \mid \theta_B))$$

and the estimated novelty score is given by

$$value_{MIX}(d; \theta_O) = \arg\max_{\lambda} l(\lambda|d, \theta_O)$$

The EM algorithm can be used to find the unique λ^* that maximizes this score.

4.1.2 Multiple Reference Topic Models

When there is more than one reference topic model, an appropriate account of the previous models must be made to compute a summarized novelty value for a document. One possibility is to compute a mixture (average) of the reference topic models, so that the problem is reduced to the single reference model case. Another possibility is to compute a novelty score for d_i using *each* previous d_j as a reference topic model θ_O, and to then combine these scores. The first method is straightforward. For the second, three obvious possibilities for combining the individual novelty scores are taking the minimum, maximum, and average. However, using the maximum distance is unreasonable, since a document would be judged as novel if it is different from a single old document d_j, even the case where it is identical to another $d_{j'}$.

With two novelty measures for a single reference model and two reasonable ways of computing a combined novelty score over multiple reference models, we have six different novelty measures, as shown in Table 1.

Basic measure	Aggregation		
	$d_1, \ldots, d_{i-1}$ averaged	d_i *vs* d_j scores combined: min	d_i *vs* d_j scores combined: average
KL	KLAvg	MinKL	AvgKL
Mixture	MixAvg	MinMix	AvgMix

Table 1: Novelty measures based on language models.

4.1.3 Comparison of Novelty Measures

We compared all six novelty measures on the subtopic retrieval task. In order to focus on the effectiveness of novelty detection alone, we considered the special task of re-ranking relevant documents, using the greedy algorithm of Figure 2 and *value* functions which are appropriate aggregations of the functions $value_{KL}$ and $value_{MIX}$. Since none of the novelty measures can be used to select the very first document, we used the query-likelihood relevance value function with Dirichlet prior smoothing; essentially all different rankings start with the same (presumably most likely relevant) document. The same query-likelihood relevance value function is also used to produce a ranking of all the relevant documents, which we use as our baseline.

We evaluated the ranking using both the S-precision and WS-precision measures. The results are shown in Figure 3. We make the following observations.

Overall, MixAvg is the best performing novelty-based ranking, followed by MinMix. Particularly at high recall levels, MixAvg is noticeably better than any of the other measures.

For both measures, the relevance baseline ranking is relatively good at low levels of subtopic recall, and relatively poor at higher levels of subtopic recall. This is intuitive, since subtopics are more likely to be duplicated later in a ranking when we will have accumulated more subtopic instances. The novelty-based ranking schemes outperform the relevance measure most consistently on the WS-precision measure. This is to be expected since the WS-measure more heavily penalizes redundancy.

The KL-based ranking schemes are generally inferior to the mixture-based ranking schemes, by both measures. They are also (perhaps surprisingly) generally inferior to the baseline relevance ranking, especially at high subtopic recall levels. The MinMix measure performs slightly better than the AvgMix measure, and similarly, the MinKL measure performs slightly better the AvgKL measure. We note that MinMix is most similar to the original MMR measure [2].

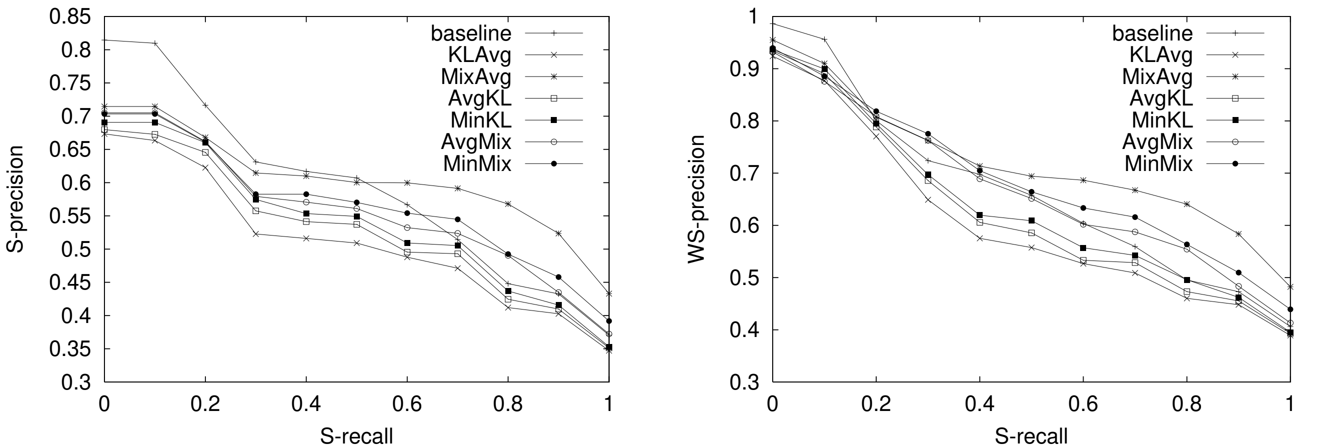

Figure 3: Comparison of the curves of S-precision (left) and WS-precision (right) versus S-recall for the six novelty measures and the baseline relevance ranking.

4.2 Combining Relevance and Novelty

We now consider how to combine novelty and relevance in a retrieval model. Based on other relevance-based retrieval experiments [15, 8], we use KL-divergence as a measure for relevance ($value_R$) and MixAvg as a measure of novelty ($value_N$). Unfortunately, a direct interpolation of these measures would not make much sense, since they are not on the same scale. We note that the MixAvg estimate of $value_N$ can be loosely interpreted as the expected percentage of novel information in the document, or the probability that a randomly chosen word from the document represents new information. Thus, we may consider two probabilities associated with a document d. One is the probability of relevance $p(Rel \mid d)$, the other is the probability that any word in the document carries some new information $p(New \mid d)$. This leads to the following general form of the scoring function

$$\begin{aligned} s(d_i; d_1, ..., d_{i-1}) &= c_1 p(Rel \mid d_i) p(New \mid d_i) \\ &+ c_2 p(Rel \mid d_i) p(\overline{New} \mid d_i) \\ &+ c_3 p(\overline{Rel} \mid d_i) p(New \mid d_i) \\ &+ c_4 p(\overline{Rel} \mid d_i) p(\overline{New} \mid d_i) \end{aligned}$$

where c_1, c_2, c_3, and c_4 are cost constants.

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

Intuitively, c_2 is the cost of user seeing a *relevant*, but *redundant* document, whereas c_3 the cost of seeing a *non-relevant* document. We will finally assume that $c_2 > 0$ (i.e., that the user cares about redundancy), which allows us to re-write the scoring function in the equivalent form

$$\begin{aligned} & s(d_i; d_1, ..., d_{i-1}) \\ &= c_3 + c_2 p(Rel \mid d) \left(1 - \frac{c_3}{c_2} - p(New \mid d)\right) \\ &\stackrel{\text{rank}}{=} p(Rel \mid d) \left(1 - \frac{c_3}{c_2} - p(New \mid d)\right) \end{aligned}$$

where $\stackrel{\text{rank}}{=}$ indicates that the two scores differ by a constant, and therefore give identical rankings. Note that a higher $p(New \mid d)$ makes the cost score better (i.e., lower). Further, when $\frac{c_3}{c_2} \geq 1$, a higher $p(Rel \mid d)$ also makes the score lower, but the amount of reduction is affected by the cost ratio $\frac{c_3}{c_2}$. This ratio indicates the relative cost of seeing a non-relevant document compared with seeing a relevant but redundant document. When the ratio is large, i.e., $c_2 \ll c_3$, the influence of $p(New \mid d)$ could be negligible. This means that when the user has low tolerance for non-relevant document, the optimal ranking would essentially be relevance-based, and not affected by the novelty of documents. When $c_3 = c_2$, we would score documents based on $p(Rel|d)p(New|d)$, which is essentially the scoring formula for generating temporal summaries proposed in [1], where $p(Rel|d)$ is referred as $p(Useful|d)$. In general, there will be a trade-off between retrieving documents with new content and avoiding retrieval of non-relevant documents.

One technical problem remains, since we do not usually have $p(Rel \mid d)$ available when we score documents with the KL-divergence function. One possible solution is to consider ranking documents based on the query likelihood, i.e., $p(q \mid d)$, which is equivalent to ranking based on the KL-divergence [7]. Since $value_R = p(q \mid d)$, we may further assume that $p(Rel \mid d)$ is *proportional* to $p(q \mid d)$. Under this assumption, the scoring function can be rewritten as

$$\begin{aligned} & s(d_i; d_1, ..., d_{i-1}) \stackrel{\text{rank}}{=} \\ & \quad value_R(\theta_i; \theta_Q)(1 - \rho - value_N(\theta_i; \theta_1, ..., \theta_{i-1})) \end{aligned}$$

where $\rho = \frac{c_3}{c_2} \geq 1$, $value_R(\theta_i; \theta_Q) = p(q \mid d_i)$ is the query likelihood, and $value_N(\theta_i; \theta_1, ..., \theta_{i-1})$ is the estimated novelty coefficient using the mixture model method. We refer to this scoring function as a *cost-based* combination of relevance and novelty.

5. EXPERIMENTS

In order to evaluate the effectiveness of the proposed method for combining novelty and relevance, we compared it with a well-tuned relevance-based ranking baseline. The baseline is the best relevance-based ranking (in terms of the subtopic coverage measure) using the original (short) queries. This baseline ranking is achieved using the Dirichlet prior ranking method [15] with smoothing parameter set to μ = 20,000. We explored two tasks: *re-ranking relevant documents* (the same task used above to evaluate novelty methods), and *ranking a mixture of relevant and non-relevant documents*. The latter task is the "real" problem of subtopic retrieval. For the sake of efficiency, the results for re-ranking a mixture of relevant and non-relevant documents are based on using a cost-based ranking scheme to re-rank the 100 top-ranked documents returned by the baseline ranking.

As a further comparison point, we also tried using pseudo-feedback on top of our simple baseline. Intuitively, since pseudo-feedback adds new terms to a query, it might be expected to increase the diversity (and hence decrease redundancy) of the documents returned as relevant. The feedback approach that we use constructs an expanded query model based on an interpolation of the original maximum-likelihood query model and a pseudo-feedback model with a weight of $\frac{1}{2}$ on each. The feedback model is estimated based on the top 100 documents (from the simple baseline results) using a mixture model approach to feedback [14] (with the background noise parameter set to 0.5.) The Dirichlet prior smoothing parameter is set to $\mu = 5,000$, which is approximately optimal for scoring the expanded query.

We varied the cost parameter ρ between 1 and 10. Note that it is unreasonable to set ρ to any value below 1, as it would mean that a larger relevance value corresponds to greater cost. As ρ becomes large, the combination relies more on relevance; with $\rho = 10$, the formula is almost completely dominated by relevance. Notice that subtopic performance can be improved by either improving relevance ranking and keeping redundancy fixed, by improving redundancy and keeping relevance fixed, or by improving both relevance and redundancy.

5.1 Re-ranking relevant documents

Figure 4 presents the results on the simpler task of re-ranking relevant documents. We show results for the cost-based method with $\rho = 5$ and $\rho = 1.5$. Combining relevance and novelty with either weighting scheme gives a consistent improvement over both baselines, across all but the lowest recall levels, and for both measures. This is in contrast to using novelty scores alone, which improved over the baseline only for higher subtopic recall levels. This is desirable behavior for a method that combines relevance (which does well at low subtopic recall levels) with novelty (which does well at high recall levels). Feedback barely improves upon the baseline retrieval method.

5.2 Ranking mixed documents

Results are presented in Table 2 for the more difficult task of ranking a mixed pool of documents, along with an "upper bound" of performance which will be discussed in Section 5.3. We see that the cost-based combination method still improves over the baseline on both measures, but only slightly, and only for larger values of ρ. Interestingly, the pseudo-feedback approach also improves slightly over the baseline method for both S-precision and WS-precision. In fact, for S-precision the improvement obtained by the feedback method is somewhat *larger* than the improvement obtained by the cost-based combination of novelty and relevance.[3]

5.3 Analysis and Discussion

It is likely that with the addition of non-relevant documents, performance gains due to improving the novelty of documents in a ranking are largely offset by corresponding performance losses due to imperfect relevance ranking. Since a relevant document is much more likely to overlap with another relevant document than is a non-relevant document, emphasizing novelty may well tend to move non-relevant documents up in the ranking. It is possible that the gains obtained by increasing the rank of novel relevant documents are largely offset by the cost of also pulling up non-relevant documents in the ranking.

[3] Graphs are not shown for these results, but the curves for all the methods track each other quite closely.

Ranking Method	Avg S-Precision		Avg WS-Precision	
baseline	0.332	—	0.468	—
cost, $\rho = 1.5$	0.305	-8.1%	0.456	-2.6%
cost, $\rho = 5$	0.339	+2.1%	0.474	+1.2%
baseline+FB	0.344	+3.6%	0.470	+0.4%
"upper bound"	0.416	+25.3%	0.516	+10.3%

Table 2: Comparison of S-precision and WS-precision, averaged across 11 S-recall levels, for the task of re-ranking a mixture of relevant and non-relevant documents, using the cost-based combination of MixAvg novelty and a KL-divergence based relevance ranking.

This hypothesis is supported by the performance of the cost-based method on the task of re-ranking relevant documents. To further test this possibility, we conducted another test. Recall that the definitions of (weighted) S-precision and S-recall are based on comparing a ranking system $\mathcal{S}$ with an optimal system $\mathcal{S}_{opt}$. One can use the same methodology to compare any two ranking systems. To simplify the discussion, let us call the system playing the part of $\mathcal{S}$ in a test of this sort the *benchmarked system* and the system playing the part of $\mathcal{S}_{opt}$ the *target system*. Define the *WS-precision (at r) of a benchmarked system $\mathcal{S}_1$ relative to a target system $\mathcal{S}_2$* as

$$\textit{WS-precision at } r \equiv \frac{\mathrm{minCost}(\mathcal{S}_2, r)}{\mathrm{minCost}(\mathcal{S}_1, r)}$$

Relative WS-precision is a measure of the *difference* in performance between $\mathcal{S}_1$ and $\mathcal{S}_2$—the lower the WS-precision, the larger the performance difference.

We took the rankings produced by the baseline retrieval system, henceforth $\mathcal{S}^{base}$, and removed all non-relevant documents, to produce rankings from a hypothetical system $\mathcal{S}^{base}_{relOnly}$. We then performed the same transformation on the cost-based ranking for $\rho = 5$, henceforth $\mathcal{S}^{cost}$, to produce rankings for the hypothetical system $\mathcal{S}^{cost}_{relOnly}$.

Our conjecture is that the cost-based method ranks relevant documents better than the baseline system, but also ranks non-relevant documents higher. Stated in terms of these hypothetical ranking systems, the conjecture is that (a) WS-precision for $\mathcal{S}^{base}$ relative to $\mathcal{S}^{base}_{relOnly}$ will be higher (i.e., indicate a smaller difference in performance) than the WS-precision for $\mathcal{S}^{cost}$ relative to $\mathcal{S}^{cost}_{relOnly}$ and (b) WS-precision for $\mathcal{S}^{base}_{relOnly}$ relative to $\mathcal{S}_{opt}$ will be lower (i.e., indicate a larger performance difference) than the WS-precision for $\mathcal{S}^{cost}_{relOnly}$ relative to $\mathcal{S}_{opt}$.

This conjecture is confirmed by experiments; the results are shown in Figure 5. For clarity, we show WS-precision at intermediate levels of S-recall, where the differences between the systems are greatest.

A final set of experiments on ranking a mixed pool of documents was based on the observation that none of the methods considered more than modestly improves performance over the original relevance baseline. For each query, we created a subtopic query, or "subquery," for each subtopic, by concatenating the original query Q with the description of the subtopic. For instance, for the sample query 392i, we created 35 subqueries, the first of which was "What are the applications of robotics in the world today? 'clean room' applications in healthcare & precision engineering." We then retrieved the top 500 documents for each subquery, using the base-

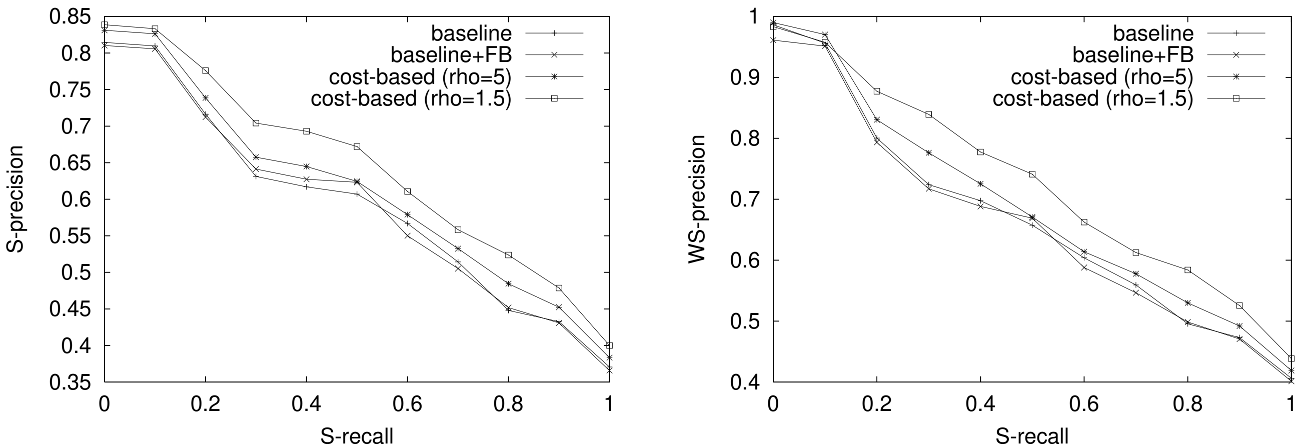

Figure 4: Comparison of the curves of S-precision (left) and WS-precision (right) versus S-recall for the task of re-ranking relevant documents, using a cost-based combination of MixAvg for novelty, and a KL-divergence measure for relevance.

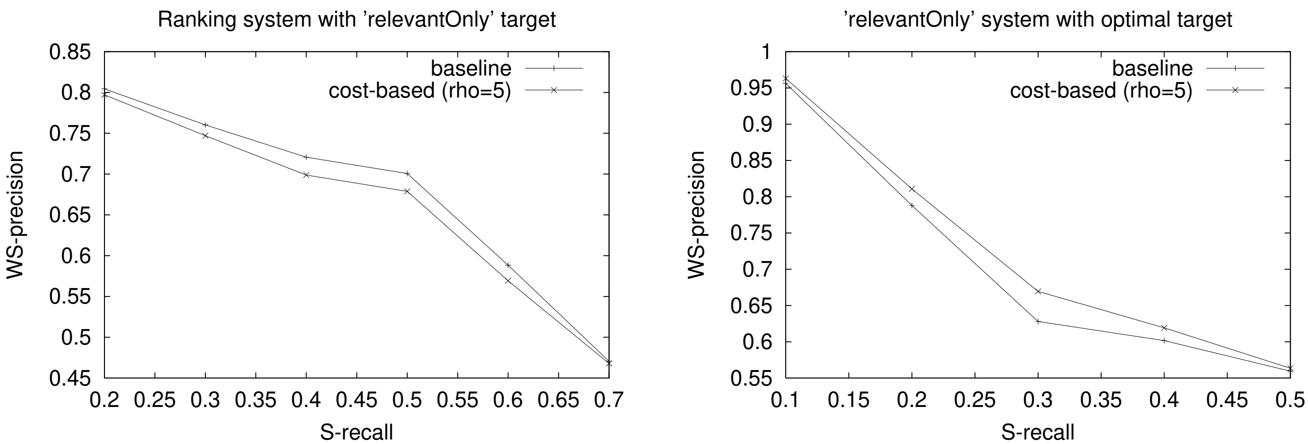

Figure 5: Comparison of relative WS-precision versus S-recall for the task of re-ranking a mixed pool of documents. On the left, WS-precision of a "real" ranking system relative to a hypothetical ranker that rejects all non-relevant documents, but otherwise does re-order documents. On the right, WS-precision of the hypothetical relevant-document-only ranking relative to the optimal ranking.

line method *with* pseudo-feedback, and placed all of the documents returned by any subquery for Q into a single pool for Q. Finally, we ran a noise-tolerant version of a greedy set-covering algorithm. This algorithm uses as a value function the expected number of new subtopics covered by a document, using subquery relevance scores to estimate the relevance of a document to a subtopic.

Unlike the MMR-style algorithms considered above, this algorithm uses an *explicit* model of the subtopics, which is acquired from the subtopic descriptions using pseudo-feedback. It is quite unreasonable to assume that this much information is available in practice, especially given that the user is unlikely to know all the subtopics in advance. However, it may be useful to consider the performance of this system as an informal upper bound on the performance of retrieval systems that must operate without any explicit model of subtopics.

The performance of this method is shown in Table 2 under the title "upper bound." Average S-precision and averaged WS-precision are improved, but by surprisingly little: S-precision is improved by about 20% over the best realistic method (the baseline with feedback), and WS-precision is improved by about 9% over the best realistic method (cost-based retrieval with $\rho = 5$).

6. CONCLUDING REMARKS

In this paper, we studied a non-traditional subtopic retrieval problem where document ranking is based on *dependent* relevance, instead of *independent* relevance, as has been assumed in most traditional retrieval methods. The subtopic retrieval problem has to do with finding documents that cover as many different subtopics as possible, which is often desirable (e.g., when the user is performing a survey on some topic). Traditional retrieval methods and evaluation metrics are insufficient for subtopic retrieval since the task requires the modeling of dependent relevance.

We proposed a new evaluation framework for subtopic retrieval, based on the metrics of S-recall (subtopic recall) and S-precision (subtopic precision). These measures generalize the traditional relevance-based recall and precision metrics, and account for the intrinsic difficulty of individual topics—a feature necessary for subtopic retrieval evaluation. We also introduced WS-precision (weighted subtopic precision), a further generation of S-precision that incorporates a cost of redundancy.

We proposed several methods for performing subtopic retrieval based on statistical language models, taking motivation from the maximal marginal relevance technique. We evaluated six novelty measures, and found that a simple mixture model is most effective. We then proposed a cost-based combination of this mixture model novelty measure with the query likelihood relevance ranking. This method was shown to slightly outperform a well-tuned relevance ranking baseline. However, the improvement is most clearly seen for ranking only relevant documents; when working on a mixed set of relevant and non-relevant documents, the improvement is quite small, slightly worse than a tuned pseudo-feedback relevance ranking of the same documents. This indicates that while both relevance and novelty/redundancy play a role in subtopic retrieval, relevance is a dominating factor in our data set.

In future work, we need to further study the interaction of relevance and redundancy, perhaps by using synthetic data to control factors such as the level of redundancy and the number of subtopics. A major deficiency in all of the MMR style approaches considered here is the *independent* treatment of relevance and novelty. As a result, there is no direct measure of relevance of the new information contained in a new document. Thus, a document formed by concatenating a seen (thus redundant) relevant document with a lot of new, but non-relevant information may be ranked high, even though it is useless to the user. It will be interesting to study how to identify and measure the relevance of the novel part of a document, which is related to the TREC novelty track [4].

ACKNOWLEDGMENTS

We thank James Allan, Jamie Callan, Jaime Carbonell, Rong Jin, and the anonymous reviewers for helpful comments on this work. This research was sponsored in part by the Advanced Research and Development Activity in Information Technology (ARDA) under its Statistical Language Modeling for Information Retrieval Research Program, contract MDA904-00-C-2106.

7. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of SIGIR 2001*, pages 10–18, 2001.

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

[3] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, July 1998.

[4] D. Harman. Overview of the trec 2002 novelty track. In *Proceedings of TREC 2002*, 2002.

[5] W. Hersh and P. Over. Trec-8 interactive track report. In E. Voorhees and D. Harman, editors, *The Seventh Text REtrieval Conference (TREC-8)*, pages 57–64, 2000. NIST Special Publication 500-246.

[6] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of ACM SIGIR 2000*, pages 41–48, 2000.

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

[8] P. Ogilvie and J. Callan. Experiments using the lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference*, pages 103–108, 2002.

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

[10] P. Over. Trec-7 interactive track report. In E. Voorhees and D. Harman, editors, *The Sixth Text REtrieval Conference (TREC-7)*, pages 65–72, 1999. NIST Special Publication 500-242.

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

[12] T. Saracevic. Relevance reconsidered. In *Proceedings of the 2nd Conference on Conceptions of Library and Information Science*, pages 201–218, 1996.

[13] H. R. Varian. Economics and search (Invited talk at SIGIR 1999). *SIGIR Forum*, 33(3), 1999.

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

[15] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'2001*, pages 334–342, Sept 2001.

[16] Y. Zhang, J. Callan, and T. Minka. Redundancy detection in adaptive filtering. In *Proceedings of SIGIR'2002*, pages 81–88, Aug 2002.