Evaluation of Filtering Current News Search Results

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder
Information Retrieval Laboratory
Illinois Institute of Technology
10 W. 35th Street
Chicago, IL 60616
{steve,ej,abdur,grossman,frieder}@ir.iit.edu

ABSTRACT

We describe an evaluation of result set filtering techniques for providing ultra-high precision in the task of presenting related news for general web queries. In this task, the negative user experience generated by retrieving non-relevant documents has a much worse impact than not retrieving relevant ones. We adapt cost-based metrics from the document filtering domain to this result filtering problem in order to explicitly examine the tradeoff between missing relevant documents and retrieving non-relevant ones. A large manual evaluation of three simple threshold filters shows that the basic approach of counting matching title terms outperforms also incorporating selected abstract terms based on part-of-speech or higher-level linguistic structures. Simultaneously, leveraging these cost-based metrics allows us to explicitly determine what other tasks would benefit from these alternative techniques.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms: Measurement, Human Factors.

Keywords: Results filtering, evaluation, news search.

1. INTRODUCTION

Evaluation of traditional information retrieval tasks focuses on the positions of relevant documents in ranked result sets, treating the false-positive error of retrieving non-relevant documents as equivalent to its counterpart of missing relevant documents. We examine a task in which false-positives are much more costly than missing relevant documents. In this task, related current news articles are retrieved for general web queries and a limited number of them are displayed above all other general web search results, as is done by many popular search services such as Google™, Yahoo!™, and AOL™. Displaying non-relevant results in this position causes a negative user experience analogous to pop-up ads or spam. In addition to this, there are several other factors that make this task significantly different from others that have been studied: As it contains only current news articles from the past week, the number of documents in the collection is very small (848 documents in our experiment) and they are constantly changing. Of the 15% of logged queries that retrieve at least one result from the current news, only a small fraction (22%) have any relevant documents. Finally, specific (high-precision) topical intent must be determined from a typical (short) web query, exacerbating problems of disambiguation and localization. Evaluation of such a task demands a perception-oriented approach that directly measures the impact of false-positive errors. This, in combination with the small fraction of queries for which there are relevant documents and the essentially unranked display of results make traditional retrieval metrics unsuitable. Rather, we apply metrics traditionally used in document filtering tasks to this retrieval problem, allowing us to directly examine the tradeoff between false-positives and missing relevant documents. Using these metrics, we evaluate some initial threshold-based result set filtering techniques that employ basic natural language processing techniques for term selection.

2. PRIOR WORK

Although there is much work devoted to improving web search effectiveness in general, it is primarily focused on the traditional retrieval problem of producing a ranked result set for a given user query. Even in navigational search evaluations where there is often only a single relevant document, metrics are defined by the rank at which relevant documents are retrieved. Several studies examine general news search and summarization tasks. Unlike these, we focus on the alternative problem of displaying a very small amount of surrogate information (titles) for short, general web queries when related news is found in a small database of current articles. Henzinger, et. al study the task of finding current news articles that supplement topics in a stream of TV broadcast news [2]. They automatically generate queries based on the broadcast and filter result sets using score thresholds to optimize precision and recall of ranked search results, also independently evaluating three filtering rules by examining their error rates at a fixed threshold. Chadrasekar and Srinivas examine a result set filtering approach to the traditional web search problem and examine traditional IR metrics for two single-term queries, finding that higher-level supertagging and noun-phrase/verb-phrase chunking provide for better filter expressions than simple part-of-speech ones [1]. Again, their precision and recall-based evaluation does not examine the tradeoff between false-positives and misses, and their techniques do not incorporate tunable thresholds. Similarly, the TREC High Precision Track focused on precision at a somewhat low level of recall (15 results), but did not address the cost of a non-relevant document in its evaluation of interactive query sessions where users issue multiple query refinements.

Document filtering tasks such as those in TDT (Topic Detection and Tracking) and the TREC Filtering Track examine the ability of systems to identify relevant documents from a stream over time based on a persistent user profile (defined by a training set of relevant documents) rather than the immediate information need of a user query. Although this makes them differ significantly from the task we address, the metrics used in their evaluations provide insight into the tradeoff between false-positives and misses that we

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

desire [3]. TDT uses a measure that allows for varying the cost of misses versus false positives (which they term false alarms). They represent the tradeoff between these costs with a DET curve. However, the standard TDT cost function assumes that the ratio of relevant to non-relevant documents is uniform across queries [4].

3. METHODOLOGY

We adopt the TDT cost function to evaluate our result-filtering task. Traditional information retrieval measures are inappropriate for measuring users' perceptions of results in this task as they do not explicitly represent the tradeoff between costs of false-positives and misses. In addition, results displayed consist of at most three documents and typically less, yielding an effectively unranked set in terms of user perception. The TDT cost function is based on the probabilities of missing relevant results, *P(miss)*, and retrieving non-relevant results, *P(fa)*, combining them by explicitly assigning a cost to each, C_{miss} *and* C_{fa}, and weighting this combination by the relative amount of relevant documents in general, *P(rel)*, as shown in Equation 1.

$$C_{\det} = C_{miss} P(miss) P(rel) + C_{fa} P(fa)(1 - P(rel))$$

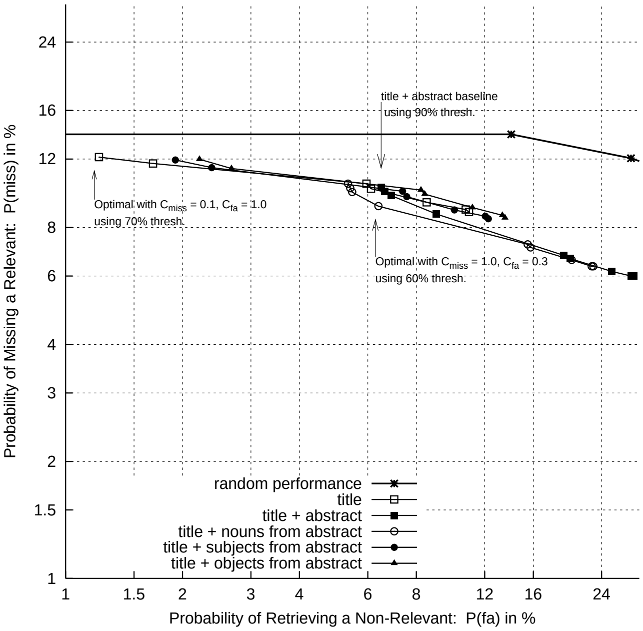
Equation 1. The TDT Cost Function

However, the detection task differs significantly from ours, resulting in several interpretive differences when applying these metrics for a perception-centric search evaluation: We treat each result from the unfiltered, ranked result sets that passes the filter and is displayed to the user (a maximum of the top three) as a "decision" rather than making decisions for each document in an entire collection. This makes our application of the metrics highly topic-weighted, as we examine averages over a much larger set of queries with many fewer decisions per query. Perhaps more importantly, it has the effect of measuring *P(miss)* and *P(fa)* from the user's perspective. In addition, the majority of our queries have no relevant results, causing the a priori probability of a document being relevant to a given query to vary drastically across queries. Therefore, we adopt the method proposed by Manmatha, et. al of calculating *P(rel)* for each query individually as we average across queries [3].

We experimented with three types of filters, each of which reject results below a minimum threshold on the number of query terms matching the title and a subset of the abstract. They differ in how they select this subset of abstract terms. Our baseline filters use title alone, or title and all abstract terms. Our second filter selects only the subset of abstract terms identified as nouns by the Brill part-of-speech tagger. Our third set of filters include all terms from phrases in the abstract identified as sentence subjects or objects by the CMU Link parser.

4. RESULTS

We randomly sampled 1,409 queries that returned at least one result from the collection of 848 documents gathered from a popular news service during the one-week period of 1/29/04-2/4/04. The average probability of a result from this set being relevant, *P(rel)*, was found to be 11%. The DET curves for the results of each filter at 10 different thresholds on percentage of matching query terms are shown in Figure 1, framed by the performance of randomly filtering varying percentages of the results. The extremities of random performance are not filtering at all (*P(miss)*=0%, *P(fa)*=85%) and filtering every document at (*P(fa)*=0%, *P(miss)*=15%). When focusing on minimizing negative user experiences from non-relevant documents, we find that simply using the title terms performs as well or better than any other techniques. If we set C_{miss}=0.1 and C_{fa}=1.0 (false alarms cost ten times as much as misses), for example, the cost for the simple title filter at a threshold of 70% or above is 0.018, while the highest threshold points of incorporating subjects or objects trail closest at 0.024 and 0.027, respectively. If we shift our focus to retrieving a larger number of relevant documents, however, by setting C_{miss}=1.0 and C_{fa}=0.3, then incorporating nouns from the abstract with a threshold of 60% begins to outperform title-only and outperforms including all terms from the abstract by 10%.



title + abstract baseline using 90% thresh.

Optimal with C_{miss} = 0.1, C_{fa} = 1.0 using 70% thresh.

Optimal with C_{miss} = 1.0, C_{fa} = 0.3 using 60% thresh.

Figure 1. DET Curves for Filtering Techniques and Random

5. CONCLUSION AND FUTURE WORK

We describe an ultra-high precision result filtering task in which users' perception is dramatically influenced by the presence of non-relevant results. We adapted a cost-based evaluation methodology to show that in this case, simple thresholding on the number of matching title terms outperforms more advanced natural language techniques. By exploiting this cost-based evaluation, however, we also find that tasks in which more non-relevant results can be afforded would benefit by selecting nouns from the abstract for use in filtering. Future work will examine more sophisticated threshold-setting and natural language techniques, as well as the incorporation of clarity metrics to determine the likelihood that a query has a relevant document.

6. REFERENCES

[1] Chadrasekar, R. and Srinivas, B. Glean: Using Syntactic Information in Document Filtering. *Inf. Process. Manage., 34*, 5 (1998), 623-640.

[2] Henzinger, M., Chang, B.-W., Milch, B. and Brin, S., Query-Free News Search. In *WWW '03*, 1-10.

[3] Ma, L., et. al., Extracting Unstructured Data from Template Generated Web Documents. In *CIKM '03*

[4] Manmatha, R., Feng, A. and Allan, J., A Critical Examination of TDT's Cost Function. In *SIGIR '02*, 403-404.