Improving Text Collection Selection with Coverage and Overlap Statistics

Thomas Hernandez
Arizona State University
Dept. of Computer Science and Engineering
Tempe, AZ 85287
th@asu.edu

Subbarao Kambhampati
Arizona State University
Dept. of Computer Science and Engineering
Tempe, AZ 85287
rao@asu.edu

ABSTRACT

In an environment of distributed text collections, the first step in the information retrieval process is to identify which of all available collections are more relevant to a given query and which should thus be accessed to answer the query. We address the challenge of collection selection when there is full or partial overlap between the available text collections, a scenario which has not been examined previously despite its real-world applications. To that end, we present COSCO, a collection selection approach which uses collection-specific coverage and overlap statistics. We describe our experimental results which show that the presented approach displays the desired behavior of retrieving more new results early on in the collection order, and performs consistently and significantly better than CORI, previously considered to be one of the best collection selection systems.

Poster recommended by WWW2005 Program Committee.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Retrieval models, Selection process*

General Terms: Management

Keywords: collection selection, collection overlap, statistics gathering

1. INTRODUCTION

With the recent emergence of meta-search engines, news meta-searchers, and bibliography search engines, it has become apparent that the challenge of retrieving relevant documents from a group of collections involves more than just searching every information source at hand and ranking the results afterwards. An effective system must first choose which collection or subset of collections to call to answer a given query. This particular process is generally referred to as *collection selection*. This is especially important because redundant or irrelevant calls can be expensive in terms of query execution cost, post-query processing (i.e. duplicate removal and results merging), network load, source load, etc. Naturally, as the number of collections increase, effective collection selection becomes essential for the performance of the overall retrieval system.

The general trend in the existing approaches for collection selection [5] is to evaluate the "goodness" of each collection based on some type of information about term, document, and/or collection frequencies. In other words, these approaches require some term frequency statistics about each collection in order to select the sources they deem relevant to the query. This general strategy works fairly well when the collections do not overlap. However, because all of these approaches fail to take into account overlap between collections when determining their collection order, they may decide to call a collection which has no – or very few – new documents (considering the documents which have already been retrieved at that moment). Take for example the case of two mirror collections. If one is deemed highly relevant, the other one would also be highly relevant, and hence both collections would be called even though calling the second one does not provide any new results.

Evidently a collection selection approach which could prevent unnecessary collection accesses would be useful and probably complementary to the existing approaches. Our motivation was thus to design a system able to order the collections such that when a collection is accessed, it is the collection which would provide the most new results. To do so, our system must be capable of making two types of predictions: *(a)* how likely a collection is to have relevant documents, and *(b)* whether a collection is useful given the ones already selected.

This paper presents our collection selection approach, called[1] COSCO, which uses information on the coverage of individual collections to predict the first point, and information on the overlap between collections to predict the second point. While it is easy to see that coverage and overlap information regarding the collections will help in the collection selection, the open issue is how to efficiently gather this information.

2. THE COSCO APPROACH

COSCO is essentially composed of an offline component which gathers statistics from collections and an online component which uses the statistics at runtime to determine the collection ranking for a new incoming query.

2.1 The Offline Component

The offline component must first obtain the appropriate coverage and overlap information from the collections for a set of training queries. Overlap between two *text* collections means that some documents are highly *similar*, as opposed to strictly identical. The complexity of computation is thus mainly affected by the two following observations. First, collection overlap is non-symmetric, in that a single result in collection C_1 could very well be highly similar to several

[1] COSCO stands for **CO**llection **S**election with **C**overage and **O**verlap Statistics

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

results in C_2. Second, document overlap is not transitive, as a document in C_1 can overlap with document d_2 in C_2 as well as with document d_3 in C_3, even though d_2 is not highly similar to d_3.

To address these inherent challenges, we compute overlap between two collections for a particular keyword query as follows: for each collection, the documents returned for the query are put into a single bag of keywords. Then, the similarity between the keyword bags is used to approximate overlap between the two collections for that particular query. Furthermore, we only compute pairwise overlaps between collections. In addition to overlap statistics, the offline component also retrieves coverage statistics, which simply refers to the number of documents a collection returns for a specific query. Both coverage and overlap statistics are collected using a list of past queries.

Next, the offline component identifies frequent item sets among the previously asked queries. Finally it computes new statistics corresponding to each item set by averaging the statistics for each query that contains the item set. The justification for the frequent item set computation is that by storing statistics with respect to item sets, we can effectively map at runtime new queries to a set of item sets for which we store statistics, and then use these to approximate statistics for the new query.

2.2 The Online Component

The online component encompasses three phases. First the incoming query must be mapped to a set of item sets for which the system has statistics. The mapping is accomplished by determining which group of item sets covers most, if not all, of the query. Second, coverage and overlap statistics for the query are computed by averaging the statistics of all mapped item sets.

Finally, using these estimated query statistics, the system determines which collections to call and in what order. The first collection selected is simply the one with highest estimated coverage. The next collections are selected by determining which one would lead to the largest remaining result set document, taking into account the estimated overlap between collections. More formally, at each step k we select collection C_l such that

$$l = \begin{cases} \text{for } k = 1: \quad \underset{i}{argmax}\left[coverage_{q_{new}}(C_i) \right] \\ \text{for } k > 1: \\ \quad \underset{i}{argmax}\left[|\mathcal{R}_{iq_{new}}| - \sum_{C_j \in \mathcal{S}} overlap_{q_{new}}(C_i, C_j) \right] \end{cases}$$

where $\mathcal{R}_{iq}$ is the bag corresponding to the union of the result documents for query q from collection C_i, and $\mathcal{S}$ is the set of already selected collections. More details of our approach are contained in [3].

3. EXPERIMENTAL RESULTS

In order to determine how well COSCO performed in an environment of overlapping text collections, we set up a collection test bed as well as a set of queries and compared the respective performances of COSCO, CORI (which is a leading approach for collection selection [5]), and an oracle-like collection selection strategy which truly knows which are the best collections to access. The collection test bed was composed of 6 real online collections (ACM Digital Library, ACM Guide, ScienceDirect, Compendex, CiteSeer, and the CS Bibliography) and 9 synthetic collections, which were created with the intent of having both a relatively large test bed and a controlled degree of overlap between the collections. In this setup, the documents were essentially publications containing title, abstract, author names, etc. The list of queries consisted of 1,062 distinct real user queries gathered by the BibFinder mediator [1, 4].

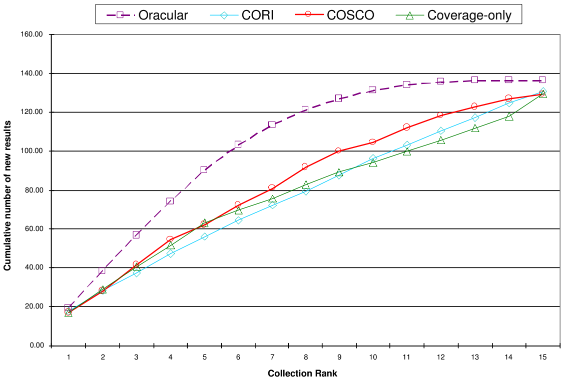

Figure 1: Performance of *Oracular*, CORI, COSCO, and a variation of our approach on the 15-collection test bed.

COSCO's offline component used 90% of the query-list to probe each collection, identify frequent item sets in the training query-list, and gather coverage and overlap statistics for the online component to use. The remaining 10% of the query-list were then used to test all three collection selection approaches. We kept track of the cumulative number of new[2] results retrieved in terms of the number of collections called, in order to analyze to what degree each approach was able to retrieve more results in the first few collections called. Figure 1 displays the *cumulative* plots of the different approaches, and illustrates how COSCO usually retrieved more results than CORI in the same number of collection calls. More experimental results are given in [3].

4. CONCLUSION AND FUTURE WORK

Probably the most interesting direction of future work follows from the fact that in our approach the relevance of the results from each collection does not guide at all the final collection ranks. Therefore an interesting extension attempts to design a collection selection system which would take into account both the content-based relevance of the documents and/or collections, as well as the overlap between the collections. This essentially considers our work as a complementary strategy to those that have been proposed in the literature, and preliminary work on the subject seems to point to a promising system.

5. REFERENCES

[1] BibFinder: A Computer Science Bibliography Mediator. `http://rakaposhi.eas.asu.edu/bibfinder`, 2004.

[2] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of ACM SIGIR Conference*, pages 21–28, 1995.

[3] T. Hernandez and S. Kambhampati. Collection selection in the presence of overlapping text collections. *Submitted to ACM SIGIR*, 2005. `http://rakaposhi.eas.asu.edu/cosco.html`

[4] Z. Nie and S. Kambhampati. A frequency-based approach for mining coverage statistics in data integration. In *Proceedings of the ICDE Conference*, 2004.

[5] A. L. Powell and J. C. French. Comparing the performance of collection selection algorithms. *ACM Transactions on Information Systems*, 21(4):412–456, 2003.

[2] A new result is one that is not highly similar to a result which has been retrieved previously.