Searching Archives of Community Knowledge

Research and Education Activities:

One of the emerging trends in web information services has been the growth in sites where people answer other peoples' questions. This includes collaborative question-answering (CQA) services such as Yahoo! Answers, as well as forums, blogs, and FAQ sites. Over time, these services build up very large archives of previous questions and their answers. These archives represent a new type of community-based knowledge that supplements and, in many cases, improves on the usual information found using standard web search. Many of the questions answered regularly using this service do not, in fact, produce any useful results in a standard ranking of web pages. The fact that people are responding to other peoples' questions means that, in contrast to web search, the questions that people ask are quite long. In these types of services, it appears that people really ask what they want to know, rather than being forced to think up a couple of keywords to attempt to capture the essence of their question. These archives of community knowledge, therefore, represent new and exciting linguistic resources that enable us to investigate approaches to dealing with longer, more detailed questions and break through the "keyword barrier" of web search.

The original goals of this project as stated in the proposal were to:

1. Develop techniques for finding semantically similar questions in query archives and evaluate by measuring the accuracy of the answers that are retrieved.
2. Use large query archives to learn transformation models that can be used in more general retrieval environments.

We have had considerable success in these areas and have produced papers at major conferences describing this work. To accomplish this, we have used a variety of resources to pursue this research, including CQA data, forums, blogs, patents, and the web. The research has also broadened somewhat over the course of the project, as described in the following paragraphs, although the contributions can all be related to these two high-level goals.

With regards to the first goal, the most important paper is Xu, Croft and Jeon (2008), where we published the results of a comprehensive series of experiments showing that using translation models to find similar questions in a CQA archive can significantly improve the effectiveness of retrieval. In our work with forum data (Seo, Croft and Smith, 2009), we showed that features indicating similarity of forum entries can be used to construct "conversations" that in turn significantly improve the retrieval of answers. Seo and Croft (2010) also described an effective estimation technique used in our probabilistic models that originated in our work with blog data.

Much of our more recent work has focused on developing more general query transformation techniques. In papers produced this year, we have described techniques

for query transformation based on anchor text archives used as a query log (Dang and Croft, 2010), techniques for query transformation based on passage analysis, (Xue, Croft and Smith, 2010), incorporating query transformations into retrieval models (Xue and Croft, 2010), techniques for ranking alternate query formulations (Dang, Bendersky and Croft, 2010), techniques for transforming patent documents into queries (Xue and Croft, 2009), query transformation techniques based on dependency parsing (Park and Croft, 2010), and techniques for transforming longer CQA queries into web queries (Xue, Huston, and Croft, 2010).

Findings:

We have produced a wide range of results over the course of the project, which are described in detail in the publications. At a high level, we have shown that query archives can be used to transform (or rewrite) long or short queries and, when combined with the right retrieval model, this will significantly improve retrieval effectiveness. There has been a lot of research published recently by the search industry showing that web query logs together with the associated click data can be used to rewrite and suggest queries. Our research, however, was the first to show how to use CQA archives, the first to show how to use anchor text "query logs" (which are publicly available), the first to develop ranking techniques for alternate formulations, the first to integrate reformulation models into a retrieval model, and the first to evaluate archive-based query rewriting using TREC collections.

To a give a sample of some of the results, the following are summaries of some of our papers produced in the last year of the project:

Query Reformulation using Anchor Text
Query reformulation techniques based on query logs have been studied as a method of capturing user intent and improving retrieval effectiveness. The evaluation of these techniques has primarily, however, focused on proprietary query logs and selected samples of queries. In this work, we describe how anchor text, which is readily available, can be an effective substitute for a query log and study the effectiveness of a range of query reformulation techniques (including log-based stemming, substitution, and expansion) using standard TREC collections. Our results show that log-based query reformulation techniques are indeed effective with standard collections, but expansion is a much safer form of query modification than word substitution. We also show that using anchor text as a simulated query log is as least as effective as a real log for these techniques.

Learning to Rank Query Reformulations
Query reformulation techniques based on query logs have recently proven to be effective for web queries. However, when initial queries have reasonably good quality, these techniques are often not reliable enough to identify the helpful reformulations among the suggested queries. In this work, we show that we can use as few as two features to rerank a list of reformulated queries, or expanded queries to be specific, generated by a log-

based query reformulation technique. Our results across five TREC collections suggest that there are consistently more useful reformulations in the first two positions in the new ranked list than there were initially, which leads to statistically significant improvements in retrieval effectiveness.

Modeling Query Reformulation as a Probabilistic Distribution of Queries
Query reformulation modifies the original query with the aim of better matching the vocabulary of the relevant documents, and consequently improving ranking effectiveness. Previous models reformulate the original query as a bag of query components, including words and phrases, but ignores how to use those components to construct new queries. In this work, a novel framework is proposed that reformulates the original query as a probabilistic distribution of queries, where each query is a reformulation of the original query. This approach considers a query as a basic unit and thus captures the dependencies of query components. A passage analysis based implementation is described, where passages containing query words are used to generate reformulated queries and estimate the corresponding probabilities. Experiments on TREC collections show that the proposed model for query reformulation significantly outperforms state-of-the-art methods.

Improving Verbose Queries using Subset Distribution
Dealing with verbose (or long) queries poses a new challenge for information retrieval. Selecting a subset of the original query (a "sub-query") has been shown to be an effective method for improving verbose queries. In this work, the distribution of sub-queries ("subset distribution") is formally modeled within a well-grounded framework. Specifically, sub-query selection is considered as a sequential labeling problem, where each query word in a verbose query is assigned a label of "keep" or "don't keep". A novel Conditional Random Field model is proposed to generate the distribution of sub-queries. This model captures the local and global dependencies between query words and directly optimizes the expected retrieval performance on a training set. The experiments, based on different retrieval models and performance measures, show that the proposed model can generate high-quality sub-query distribution and can significantly outperform state-of-the-art techniques.

A final piece of evidence about the significance of this work is that a number of follow-up or related research efforts have been started based on our research (for example, Google Scholar lists 25 citations for the 2008 Xue, Croft and Jeon paper).

This work is supported in part by the Center for Intelligent Information Retrieval (CIIR) and in part by the National Science Foundation (NSF III COR - 0711348).