1 Activities and Findings

1.1 Project Activities and Findings

1.1.1 Research and education activities

Over the course of this project, we have made major progress towards the goal of developing representations and models for long queries that will significantly improve the effectiveness of these queries in a variety of settings. We have published 15 papers in the major IR conferences (including two that were awarded honorable mentions for best paper), one journal paper in the final stages, and graduated two Ph.D. students (Bendersky and Xue) whose theses focused on the long query research. Both students were highly sought after by the search industry for their expertise in long queries and query models. According to Google Scholar, there have already been more than 60 citations to this research (as of August 16, 2012), with the following papers being the most highly cited:

- Bendersky, M., Metzler, D. and Croft, W. B. , "Learning Concept Importance Using a Weighted Dependence Model," Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010). **31 citations**.
- Bendersky, M., Croft, W. B. and Diao, Y., "Quality-Biased Ranking of Web Documents," Proceedings of the Fourth International Conference on Web Search and Data Mining (WSDM 2011). **10 citations**.
- Bendersky, M., Croft, W. B. and Smith, D., "Structural Annotation of Search Queries Using Pseudo-Relevance Feedback," Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2010), pp. 1537-1540. **6 citations**.
- Dang, V., Bendersky, M. and Croft, W. B. , "Learning to Rank Query Reformulations," Proceedings of the 33rd Annual ACM SIGIR Conference (SIGIR 2010). **5 citations**.
- Bendersky, M., Metzler, D. and Croft, W. B. , "Parameterized Concept Weighting in Verbose Queries," Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'11). **5 citations**.

During the project, our group also proposed and ran, in collaboration with Microsoft Research, the Query Representation and Understanding Workshop at the SIGIR Conference in 2010 and 2011. This workshop was very successful and attracted a wide range of attendees from academia and industry.

In the final year of the project, Bendersky and Xue both focused on finishing up their research and produced their theses and papers that recently appeared at SIGIR 2012.

Bendersky's research has involved developing retrieval models that incorporate dependencies between the terms in the query, which he has shown are a crucial part of effective long query retrieval. In his latest work, he advances this query representation one step further, and proposes a retrieval framework that models higher-order term dependencies, i.e., dependencies between arbitrary query concepts rather than just query terms. In order to model higher-order term

dependencies, he represents a query using a *hypergraph* structure -- a generalization of a graph, where a (hyper)edge connects an arbitrary subset of vertices. A vertex in a query hypergraph corresponds to an individual query concept, and a dependency between a subset of these vertices is modeled through a hyperedge. An extensive empirical evaluation using both newswire and web corpora has demonstrated that query representation using hypergraphs is highly beneficial for verbose natural language queries. For these queries, query hypergraphs significantly improve the retrieval effectiveness of several state-of-the-art models that do not employ higher-order term dependencies.

Bendersky also did another important paper dealing with using multiple sources of information in query formulation. Most standard information retrieval models use a single source of information (e.g., the retrieval corpus) for query formulation tasks such as term and phrase weighting and query expansion. In contrast, in this paper, he presents a unified framework that automatically optimizes the combination of information sources used for effective query formulation. The proposed framework produces fully weighted and expanded queries that are both more effective and more compact than those produced by the current state-of-the-art query expansion and weighting methods. He conducted an empirical evaluation of this framework for both newswire and web corpora that showed that, in all cases, the combination of multiple information sources for query formulation is more effective than using any single source. The proposed query formulations are especially advantageous for large scale web corpora, where they also reduce the number of terms required for effective query expansion, and improve the diversity of the retrieved results.

Xue's research has looked at the problem of reformulating long or verbose queries to make them more effective. Processing these longer queries usually requires a series of query operations, which results in multiple sequences of reformulated queries. However, previous query representations, either the "bag of words" method or Xue's "query distribution" method (developed in an earlier paper), cannot effectively model these query sequences, since they ignore the relationships between two queries. In his latest paper, a reformulation tree framework is proposed to organize multiple sequences of reformulated queries as a tree structure, where each path of the tree corresponds to a sequence of reformulated queries. Specifically, a two-level reformulation tree is implemented for verbose queries. This tree effectively combines two query operations, i.e., subset selection and query substitution, within the same framework. Furthermore, a weight estimation approach is proposed to assign weights to each node of the reformulation tree by considering the relationships with other nodes and directly optimizing retrieval performance. Experiments on TREC collections show that this reformulation tree based representation significantly outperforms state-of-the-art techniques.

Xue also produced a journal paper summarizing his work on the query distribution representation. 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 typically generate words and phrases related to the original query, but do not consider how these words and phrases would fit together in actual queries. In his paper, a novel framework is proposed that models reformulation as a distribution of actual queries, where each query is a variation of the original query. This approach considers an actual query as the basic unit and thus captures important query-level dependencies between words and phrases. An implementation of

this framework that only uses publicly available resources is proposed, which makes fair comparisons with other methods using TREC collections possible. Specifically, this implementation consists of a query generation step that analyzes the passages containing query words to generate reformulated queries and a probability estimation step that learns a distribution for reformulated queries by optimizing the retrieval performance. Experiments on TREC collections show that the proposed model can significantly outperform previous reformulation models.

Another significant piece of research is appearing in a paper this year by C.J. Lee. This research deals with the effective generation of long queries through people selecting passages in documents. People browsing the web may often see text passages in a web page that describe a topic of interest. Formulating a query from that text can be difficult, however, and an effective search is not guaranteed. In this paper, to address this problem, a learning-based approach is proposed that generates effective queries from the content of an arbitrary web page. Specifically, the approach extracts and selects representative chunks (noun phrases or named entities) from the content (a text segment) using a rich set of features. Experiments show that the selected chunks can be effectively used to generate queries both in a TREC environment, where weights and query structure can be directly incorporated, and with a ``black-box" web search engine, where query structure is more limited.

1.1.2 Findings

The research carried out in this project has been instrumental in helping to establish longer queries as an important area for research and development in information retrieval. This type of research is now being addressed in a variety of papers dealing with “verbose” keyword queries, natural language queries, CQA queries, queries generated from documents, “long-tail” queries, etc. The workshops we conducted as part of the project played an important role in defining the issues that need to be addressed in this research.

In terms of the specific research carried out with this funding, in a series of well-received papers we have established the parameterized Markov Random Field model as the state-of-the-art approach for representing the structures and weights required to do effective retrieval with long queries. We showed that this model can effectively combine various definitions of query concepts and query expansion, and can incorporate information from multiple sources. The hypergraph representation based on this approach allows us to capture dependencies between concepts, which is primarily important for longer queries.

We have also showed that query reformulation is a key idea for processing longer queries. We developed models for selecting subsets of long queries and other query transformations, such as segmentations or term replacement. We showed that these models can be trained using corpus data and other sources that do not rely on the availability of large query logs. A key idea in our approach to reformulation is to represent a query as a distribution of reformulations, rather than as a bag-of-words or a single “best” reformulation. The distribution representation captures important information about the context of query words and was shown to have significant effectiveness advantages.

In our other work in this project, we have shown that features derived from dependency parsing can be used to improve the effectiveness of longer queries. There is still much work to be done here, but this was one of the first promising results incorporated parsing in IR. We have developed effective methods for generating queries from documents, including user-selected passages or entire documents such as patent applications. These methods focus on identifying the important concepts and sections of documents, and weighting them appropriately.

Many of the ideas from this research are being used and refined in other projects and proposals, both at UMass and in collaboration with other universities. Our ideas are also being studied and used in search companies through internships, collaborations, and hiring (two Ph.D.s funded by this project were recently hired by industry).