Query Representation and Understanding Workshop

W. Bruce Croft[1] Michael Bendersky[1]
[1]University of Massachusetts Amherst
croft,bemike@cs.umass.edu

Hang Li[2] Gu Xu[2]
[2]Microsoft Research Asia
hangli,guxu@microsoft.com

Abstract

This report summarizes the events of the SIGIR 2010 workshop on Query Representation and Understanding, which was held on July 23rd, 2010 in Geneva, Switzerland.

1 Workshop Overview

The Query Representation and Understanding Workshop, which was held as a part of the SIGIR 2010 conference, had the goal of bringing together the different strands of research on query understanding in order to increase the dialogue between researchers working in this relatively new area, and to develop some common themes and directions, including definitions of tasks and evaluation methodology. The workshop provided a much-needed forum for the researchers to identify the issues and the challenges in this research area, to share their latest research results, to express a diverse range of opinions about this topic, and to discuss future directions.

The workshop program was made up of three main parts: invited talks, a poster session, and a discussion session. The *invited talks* included ten 20-minute talks by both academic and industrial researchers, and gave the participants a sense of different aspects of query understanding, and what are the current state of the art results in this research area. In the *poster session*, eight short (up to four pages) accepted papers were presented in a poster form, as well as 2-minute "highlight" presentations. The *discussion session*, led by Bruce Croft and Hang Li, focused on summarizing the various issues related to query representation and understanding research, including a rigorous definition of the task, modeling for the task, evaluation, implications for IR, and future research directions.

Overall, the workshop was very successful. It attracted more than 50 registered participants. All the invited speakers and at least one of the authors of the accepted papers attended the workshop and presented their work. The presentations generated good discussions and exchange of ideas, and feedback from the participants was very positive. The full proceedings of the workshop, as well the slides from most of the invited talks are available online on the workshop website .

2 Invited Talks

The invited talks were divided into three parts and contained ten 20-minute talks from both academic and industrial researchers who have done significant work in the area of query representation and

understanding. All the talks were well attended and highly engaging: almost all the talks went over the 20 minute limit due to audience questions. In this section, we briefly describe the content of these invited talks.

1. *Fuchun Peng (Microsoft Bing): "Concepts Identification from Queries and Its Application for Search Relevance"*. In this talk, Fuchun talked about his work on automatic query segmentation and its importance for improving the query rewriting algorithms and the relevance of the results returned by the web search engines.
2. *Rosie Jones (Akamai): "Searching for Myself"*. Rosie talked about the importance of investigating long query sequences, or sessions, instead of focusing on individual queries. User sessions may give us patterns of a user's interests, as well as clues about who the user is, and how he or she is feeling.
3. *Cheng Xiang Zhai (University of Illinois at Urbana-Champaign) "Putting Query Representation and Understanding in Context: A Decision-Theoretic Framework for Optimal Interactive Retrieval through Dynamic User Modeling"*. In this talk, Cheng argued in favor of a retrieval system that takes into account the context in which the user performs the retrieval, not just the user's query. He presented a general Bayesian decision-theoretic framework for incorporating all kinds of context information to model a user's information need dynamically as the user interacts with a retrieval system.
4. *Jian-Yun Nie (University of Montreal): "Integrating Term Dependencies According to Their Utility"*. In his talk, Jian-Yun proposed a model for automatically weighting term dependencies in the query, based on their strength and utility for retrieval. The empirical results demonstrated by this model showed that it outperforms the existing ones on almost all the collections, which demonstrates the necessity to integrate term dependencies in a variable manner, according to their utility for IR.
5. *Fernando Diaz (Yahoo! Research): "Intent Triage: Quantifying the Severity of Poor Performance on Intent Classes"*. Fernando discussed the influence of the query's intent class on the performance of the retrieval system. He introduced several approaches for modeling this influence, which promote "severity-based" system design. Such design can be used to ensure, for instance, that the retrieval system performs well for queries that occur during emergency crises.
6. *Eugene Agichtein (Emory University): "Inferring User Intent from Interactions with the Search Results"*. In his talk, Eugene argued for using searcher interaction data, which is becoming increasingly available, for improving our understanding of the users information needs. Extracting meaningful signals from this data would enable a search engine to accurately infer user intent for tasks such as real-time result reranking, dynamic result presentation, and contextualized query suggestion. His talk focused on the recent progress of his research group in modeling and exploiting client-side searcher interaction data for intent inference.
7. *Patrick Pantel (Microsoft Research) "Entity Extraction for Query Interpretation"*. Patrick proposed a general information extraction framework, showing large gains in entity extraction by combining state-of-the-art distributional and pattern-based extractors with a large set of features from a 600 million document crawl of the web, one year of query logs, and a snapshot of Wikipedia. While some of the success of this framework is based on proprietary query log and web-crawl data, Patrick showed a detailed analysis of feature correlations and interactions that demonstrates that while the query log and web-crawl features yield the highest gains, easily accessible Wikipedia features can also improve the performance of entity extraction over current state-of-the-art systems.

8. *Donald Metzler (University of Southern California) "Specialized Query Understanding".* In his talk, Don argued that two extreme approaches dominate the query representation and understanding research: one is overly general ("one size fits all"), while the second one is too specialized (each query in its own class). As an alternative, he proposed a challenge to develop a robust, fully automatic approach to specialized query understanding, which resides somewhere in between these two extremes.
9. *Michael Bendersky (University of Massachusetts Amherst) "Representing Queries as Structures".* In his talk, Michael formulated a retrieval framework that represents queries as structures of concepts. He demonstrated that such formulation allows creating rich and realistic query representations, and showed how the structural query representation can serve as a basis for both existing and novel retrieval models.
10. *Gu Xu (Microsoft Research Asia) "Enrich Query Representation by Query Understanding".* In his talk, Gu argued that the query understanding and representation can be conducted at different levels or granularities of semantics, i.e. word level, sense level, topic level and structure level. To support this view, he presented two state-of-the-art methods: one method for the structure-level query representation (named-entity recognition) and another for the sense-level query representation (query refinement).

3 Short Papers

Posters associated with the short papers were on display starting on the morning of the workshop. A brief presentation session was held to give an opportunity to the authors of the papers to pique the interest of the workshop attendees in their work by a 2-minute "elevator-pitch". This session was followed by a coffee-break during which the attendees browsed the posters and held discussions with the authors. The list of the short papers that were presented in this session is as follows. The full text of these papers can be found in the workshop proceedings .

1. Xiaofei Zhu, Jiafeng Guo, Xueqi Cheng (ICT, CAS, Beijing, P.R. China): *"Recommending Diverse and Relevant Queries with A Manifold Ranking Based Approach"*
2. Wei-Yen Day, Pu-Jen Cheng (National Taiwan University): *"Visualizing Image Query Senses by Social Tags"*
3. Xiaobing Xue, W. Bruce Croft (UMass Amherst): *"Representing Queries as Distributions"*
4. Grzegorz Chrupala, Georgiana Dinu, Benjamin Roth (Saarland University): *"Enriched syntax-based meaning representation for answer extraction"*
5. Maarten Van der Heijden, Max Hinne, Suzan Verberne, Eduard Hoenkamp, Theo van der Weide, Wessel Kraaij (Radboud University Nijmegen): *"When is a query a question? Reconstructing wh-requests from ad hoc-queries"*
6. Liliana Calderon-Benavides (UPF), Cristina Gonzalez-Caro (UPF), Ricardo Baeza-Yates (Yahoo! Research): *"Towards a Deeper Understanding of the User's Query Intent"*
7. Sumio Fujita (Yahoo! Japan Corporation), Tatsuya Uchiyama (Yahoo! Japan Corporation), Georges Dupret (Yahoo! Labs), Ricardo Baeza-Yates (Yahoo! Research): *"Search Facet Creation from Click Logs"*
8. Kevyn Collins-Thompson (Microsoft Research), Joshua Dillon (Georgia Institute of Technology): *"Controlling the search for expanded query representations by constrained optimization in latent variable space"*

4 Discussion Session

A discussion session was held at the end of the day, led by Bruce Croft and Hang Li. The goal of the session was to summarize important issues and themes that had come up during the workshop, and to consider research directions for the area. The following is an overview of the discussion focused on three areas: terminology, research directions, and methodology.

4.1 Terminology

In the research literature related to this area, a number of different terms have been used to refer to the same things. At the workshop, however, there was general agreement about how these terms related to each other. Some of the main points were:

- *Query intent* (or *search intent*) is the same thing as *information need*. The notion of an information need or problem underlying a query has been discussed in the IR literature for many years, and it was generally agreed that query intent is another way of referring to the same idea.
- *Query representation* involves modeling the intent or need. *Query understanding* refers to the process of identifying the underlying intent or need based on a particular representation.
- *Intent classes, intent dimensions*, and *query classes* are all terms used to talk about the many different types of information needs and problems.
- *Query rewriting*, *query transformation*, *query refinement*, and *query alteration* are names given to the process of changing the original query to better represent the underlying intent (and consequently improve ranking).
- *Query expansion, substitution, reduction, segmentation*, etc., are some of the techniques or steps used in the query transformation process.

Another issue that came up was defining a *query* itself. Most research in this area assumes the query to be the string entered by a user. The process of transformation or rewriting, however, can produce many different representations of the query. An important point is that we are primarily focusing on *explicit* queries, but *implicit* queries exist in a number of situations where there is clearly an underlying information need but the user does not enter a string (in browsing applications, for example).

4.2 Research Directions

The importance of query understanding can be illustrated by the following example. In one snapshot of the search log of a web search engine, there were 140 different queries which actually represent the same search intent, including ‘distance between sun and earth’, ‘how far is sun from earth’, ‘sun earth distance’, ‘distance between earth and sun’, etc. Ideally, we would like to have the system return the same result for these queries. However, current web search engines are not guaranteed to perform this way. One of the significant challenges to IR research is how to use better query understanding as well as document understanding to match the query against the documents at the semantic level, rather than the term level.

Again, by query understanding we mean the process of generating a representation which characterizes a user’s search intent. Query understanding, together with document ranking and indexing, emerge as the most basic components of information retrieval systems. Query

understanding can be potentially useful for improving general search relevance, user experience, vertical search, and helping users to accomplish tasks.

A query representation may contain a variety of information, including: whether the query is informational or navigational, the semantic classes or topics of the query, time or location sensitiveness of the query, similar queries (representations) to the current query, key phrases in the query, text segments in the query, named entities and their attributes in the query, whether the query can be personalized, whether the query has commercial intent, etc.

Existing work on query understanding can be viewed as a development of the techniques for producing reliable query representations. This mainly falls into four groups: *query transformation* (rewriting, refinement, or alteration), *query classification*, *query parsing*, and *exploiting context/user information*.

The invited talks in this workshop can also be classified into these four groups.

- *Query transformation*: M. Bendersky and G. Xu;
- *Query classification*: D. Metzler and F. Diaz;
- *Query parsing*: F. Peng, J. Nie, G. Xu, P. Pantel;
- *Exploiting context and user information*: R. Jones, C. Zhai, and E. Agichtein.

There are still many open questions and research topics for query understanding. For example,

- How to develop a unified and general framework for query understanding?
- How to formally define a query representation?
- How to develop new system architectures for query understanding?
- How to combine query understanding with other components in information retrieval systems?
- How to conduct evaluation of query understanding?
- How to make effective use of both human knowledge and machine learning in query understanding?

There was considerable discussion about defining common tasks that may help to focus query understanding research. The tasks discussed included long query relevance (e.g., creating collections of long queries for evaluations in future TREC conferences), query term weighting and reduction (e.g., query-dependent stopword removal/retention techniques), similar query finding (e.g., for the purpose of query suggestions and query reformulations), query classification, named entity recognition in queries, and context-aware search. Needless to say, the key for carrying out such research is the development and the release of the appropriate test collections, which would be available to the entire research community.

4.3 Methodology

As with any IR task, making progress on query representation and understanding requires researchers to agree on evaluation methodologies and metrics so that comparisons can be made between competing approaches. Although quite a few papers on this topic have appeared in conferences and journals, there has been little consensus or discussion to date on tasks and test collections. The previous section described our discussion on tasks; here we focus on methodology and resources. Some of the most important issues discussed were:

- *TREC-style vs. "black-box" evaluations.* Many of the evaluations reported in the query representation and understanding literature have been done "in-house" using queries collected just for those experiments and then discarded, and web search engines used as a "black box", which means that there is little control over further query processing done by the search engine. While such experiments can produce valid results, considerable care needs to be taken in the experimental design. In addition, if the queries themselves and associated data are not archived and shared with other researchers, comparisons are made more difficult. Most researchers agree that a TREC-style evaluation with a shared collection of queries, documents, relevance judgments and other related data would be preferable, but there are considerable practical difficulties with accomplishing this type of sharing of industry data. This brings up the issue of the important resources for this type of research.
- *Resources*. A number of resources are available to the academic community, but they tend to be much smaller and contain fewer types of data than industry resources. The main resources, both existing and desired, that were discussed at the workshop were:
    - *Document collections*. There was general agreement that the TREC ClueWeb collection is an acceptable resource for web-scale experiments.
    - *Query collections*. This was the area where the workshop participants felt that the most progress could be made. The idea is to have academics and industry collaborate on defining collections of queries of different types, such as long queries, location queries, product queries, etc. These collections would correspond to important research tasks and industry people would verify that these queries were an interesting sample from their perspective. Associated data such as relevance judgments or click data would then be gathered for those queries with respect to the documents in the ClueWeb collection. Our hope was that by defining specific sets of queries, there would be less privacy concerns for search companies and some of their data could be made available. One issue of concern was the size of query sets that would be needed to be a realistic sample. An interesting idea that was mentioned is that we will also need a "background" query set to ensure that any technique developed for a specific query collection does not negatively affect performance for other queries.
    - *Query logs*. Getting access to this type of data has been an ongoing major issue for academics. We discussed a number of approaches that are being used to address this currently, including the existing MSN/AOL/KDD logs, logs constructed from anchor text, logs from other applications (such as Wikipedia), and logs from more restricted environments (such as an academic library).
    - *N-grams, etc*. Google and Microsoft have provided access to n-gram statistics from their web collections, which is a valuable resource for query processing research. There was some discussion of whether other statistical data such as entity frequencies and query n-grams could also be provided.

5 The Future

One final topic that was discussed was possible future workshops in the area of query representation and understanding. Although the consensus was that there was no need to repeat the same type of workshop, the participants had considerable interest in a workshop that would focus on some specific task or transformation step. The organizers plan to work with people in the community to develop this idea further.