SIGIR WORKSHOP REPORT

Predicting Query Difficulty - Methods and Applications

David Carmel and Elad Yom-Tov
IBM Research in Haifa, Israel
{*carmel,yomtov*}*@il.ibm.com*

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

1 Introduction

Estimation of query difficulty is an attempt to quantify the quality of results returned by the search system for a query. Ideally, a system that can predict difficult queries can adapt parameters or change algorithms to suit the query. Such systems could give feedback to the user, for example, by reporting confidence scores for results, and report to the system administrator regarding topics that are of increasing interest to users but are not answered well by the system.

The prediction of query difficulty has recently been recognized by the IR community as an important capability for IR systems. The Reliable Information Access (RIA) workshop, organized by NIST, investigated the reasons for system variance in performance by performing failure analysis of several state-of-the-art IR systems. One of the conclusions of that workshop was that if a system can realize the problems associated with a given topic, then current IR techniques are able to improve results significantly. This suggests that systems can improve performance by discovering which techniques to apply to which topics.

In the Robust track of TREC 2004, systems have been asked to rank the topics by predicted difficulty, with the goal of eventually being able to use such predictions to do topic-specific processing. System performance was measured by comparing the ranking of topics based on their actual precision to the ranking of topics based on their predicted precision. Prediction methods suggested by the participants have included measuring query difficulty based on the system's score of the top results, analyzing the ambiguity of the query terms, and by learning a predictor using TREC topics and their associate relevance sets as training data. The track results clearly demonstrated that measuring query difficulty is still intrinsically difficult.

The SIGIR workshop on Predicting Query Difficulty was held in August 19th, in Salvador, Brazil. The goal of the workshop was to bring together researchers and practitioners interested in query prediction to discuss and define the most relevant topics in this area, present recent results, and propose future directions for research. The topics for discussion included:

- Identifying the reasons that cause a specific query to become difficult for a given system, or for all systems
- Prediction methods for query difficulty
- Classification of queries and failure modes, with an eye toward predicting difficulty and suggesting solutions
- Evaluation methodology for query prediction
- Potential applications for query prediction
- Tools and techniques for analysis of retrieval results and failure modes

The workshop program and papers are available at `http://www.haifa.il.ibm.com/sigir05-qp/`.

2 Towards Query Specific Customization

The opening session of the workshop consisted of a keynote talk given by Donna Harman from NIST, titled "Toward query specific customization of IR systems". Donna began with the observation that retrieval systems did improve their effectiveness over several years of TREC experiments, but at a certain point they reached a plateau. She noted that effectiveness on individual queries varies widely both between queries and between systems on the same query, and suggested that researchers and system designers need to pay attention to the individual queries, and not just to the average performance, to make further improvements in overall retrieval effectiveness.

To illustrate this, she next presented a detailed overview of the analyses done as part of the 2003 Reliable Information Access (RIA) workshop. At RIA, the participants conducted detailed failure analyses of seven different retrieval systems on 45 TREC topics. These failure analyses yielded some preliminary conclusions about different failure modes and how they occur across different systems (Buckley, 2004; Harman and Buckley, 2004). In the spirit of the workshop, rather than dwell on these overall findings Donna instead walked through four specific topics, showing exactly how systems fail differently in each topic, and how these failures might (or might not) be detectable. These analyses and further data from the RIA workshop is available at `http://ir.nist.gov/ria/`.

Donna's conclusions were perhaps surprising to some participants. Instead of looking for a holy-grail solution to predicting query difficulty, she suggested that the failure modes of queries could be identified automatically. Given classified queries, it would then be worthwhile taking a closer look at the many results in IR which at face value are intriguing but experimentally have been shown to offer modest or mixed improvements. It may be the case that careful and detailed topic-by-topic analysis can show that those techniques are in fact effective solutions for classes of queries that correspond to the failure modes we see in practice.

3 Prediction Methods

The first session of the workshop dealt with several prediction methods, and features that can be used for prediction. K. L. Kwok from the City University of New York simplified the task to identifying weakest and strongest queries only. He showed that employing a combination of query features, including term inverse document frequencies and a distribution of average term frequency values, can predict correctly about 1/3 to 1/2 of the weakest and strongest 6 topics among 50 topics. Prediction difficulty of longer queries can return better results, but it seems less consistent. During his talk, he also discussed using the web to expand poorly-performing queries. He concluded that the results of prediction are still not sufficiently accurate to help improving ad-hoc retrieval.

Josiane Mothe from IRIT analyzed 16 different linguistic features of the query, looking for significant correlations between these features and the average recall and precision scores obtained by several runs of TREC systems. Each of these features can be viewed as a clue to a linguistically specific characteristic of the query, either morphological, syntactical or semantical. Two of these features (syntactic links span and polysemy value) are shown to have a significant impact on either recall or precision scores. Although the correlation values are not very high, they indicate a promising link between some query linguistic characteristics and query difficulty.

Jens Grivolla from the University of Avignon presented his work on automatic classification of queries by expected retrieval performance. A predictor was trained based on query linguistic features, as well as a set of features extracted from the retrieved set (such as the top score, the range of top-k scores, the entropy between top results, etc.). Using these features, he trained decision tree and SVM-based predictors for the effectiveness of TREC-8 participants. The experiments show that for certain systems the trained predictor is quite reliable (over 80% accuracy for several systems) while for other systems the trained predictor completely fails. Jens estimated that more training data can improve prediction significantly.

Ben He from the University of Glasgow presented their work on predicting query performance in intranet search. In this work, two predictors were evaluated, one that is based on the average inverse collection term frequency of the query terms, and one based on the query scope that measures the specificity of a query according to the ratio between the number of documents in collection and the number of documents that contain at least one of the query terms. The experimental results show that the predictors are very effective for 1 and 2-term queries, but prediction quality is decreased with the query length.

4 Prediction Applications

The remainder of the second session concentrated on how prediction of query difficulty can be used by search systems. Padmini Srinivasan from the University of Iowa talked about predicting performance for gene queries. This work explored a variety of features, most of them aim at estimating the homogeneity of the retrieved document set. Features that target specific kinds of ambiguities did not exhibit interesting correlations with the average precision score and hence were not useful. They found that a simple linear regression model built from a single feature: the number of unique RN terms (Registry Number terms, which are the chemical terms assigned to documents by the Chemical Abstracts Service) divided by the total number of RN terms in the retrieved set, is most effective at predicting the score.

Finally, Elad Yom-Tov from IBM Research presented the work on metasearch and federation using query difficulty predictor. When a query is executed, the ranked list of documents is returned from each dataset and search engine combination and the prediction of query difficulty is computed for each. The predicted difficulty is used for weighting the scores of the results and the final ranking is built by merging the lists using these weighted scores. The fusion method sets a query-by-query weight for each search engine and document collection pair, hence, expected poor results contribute less to the final merged list. Experiments with several desktop search engines searching over TREC data, and a single search engine searching over several sub-collections of TREC, showed significant improvement over state-of the-art fusion techniques both for metasearch and federation.

5 Panel Discusion

The panel brought together Ellen Voorhees (NIST), Andrei Broder (IBM Research), and Abdur Chowdhury (AOL) for a discussion about query difficulty. Ellen focused on the need for a useful and efficient evaluation tools for query prediction, and discussed the evaluation measures considered by NIST for the query prediction task of the Robust track of TREC.

Andrei and Abdur discussed the implications of difficult queries for commercial search engines and intranet search. In the intranet domain, Andrei suggested that for the most part difficult queries are the result of insufficient content, rather than a failure of the search engine. Thus, missing content queries (queries with no relevant data in the collection) will fail to be answered by any search engine. Andrei pointed out that automatic identification of missing content queries will help system administrators to improve their data.

Abdur showed the results of research performed at AOL to determine the accuracy of several internet search engines. His research demonstrated that while the overall accuracy of the search engines is relatively high on average, the distribution of the accuracy over the queries is relatively wide, and thus many queries still have low precision. Therefore, there is an acute need for commercial search engines to improve the retrieval of these "difficult queries".

6 Conclusions

One of the main goals of the workshop was to promote the discussion on prediction of query difficulty in the IR community and this goal was definitely achieved. Twenty-seven attendees from industry as well as academia,

participated the workshop and demonstrated the great amount of interest in this subject. However, although some progress has been demonstrated during the workshop, we are still lacking a deep understanding of the notion of query difficulty and a mature methodology for predicting query difficulty. Furthermore it is not yet clear how prediction can be used successfully in improving search system effectiveness. From a research perspective, there are still many important issues to address. First, a deeper understanding of why some queries fail is necessary. In addition, it is important to look for features that indicate query difficulty as well as to develop good evaluation methods for query prediction. Although the latter might depend on the specific application of the predictor, a widely adopted measure will make it possible to obtain a rigorous comparison of the different prediction algorithms. Several measures were proposed at the workshop, but none of them fully addresses the requirements from a truly efficient evaluation measure. From the keynote, we can see that we should perhaps not look so hard at predicting the overall effectiveness of a query, but instead try to identify how that query might lead the system to a known type of search failure. It was exciting to see that there are a number of good applications for effective query difficulty prediction beyond improving average precision, but they all require progress on the underlying technology before they can succeed.

Acknowledgments

We would like to thank Doug Oard, the SIGIR workshops chair, for his support and his helpful advice, to the PC members of the workshop for their helpful reviews, and especially to the workshop participants for making this workshop a success.

References

Buckley C (2004). Why Current IR Engines Fail. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), Sheffield, UK, pp. 584–585.

Harman D and Buckley C (2004). The NRRC Reliable Information Access (RIA) Workshop. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), Sheffield, UK, pp. 528–529.