

# Predicting the Effectiveness of Queries and Retrieval Systems

CLAUDIA HAUFF

# Chapter 6

## Conclusions

In this thesis we have investigated the prediction of query and retrieval system effectiveness. As we introduced the topic we clearly identified its pertinent evaluation aspects (Figure 1.1) and set the focus on two aspects in particular, namely predicting the effectiveness of queries for a particular system (EA2) and predicting the relative effectiveness of systems (EA4).

The motivation for our research efforts stems primarily from the enormous benefits originating from successfully predicting the quality of a query or a system. Accurate predictions enable the employment of adaptive retrieval components which would have a considerable positive effect on the user experience. Furthermore, if we would achieve sufficiently accurate predictions of the quality of retrieval systems, the cost of evaluation would be significantly reduced.

We have conducted our research along four lines: the pre-retrieval and post-retrieval prediction of query effectiveness, the contrast between the evaluation of predictors and their effect in practice, and, lastly, the prediction of system effectiveness.

### 6.1 Research Themes

#### 6.1.1 Pre-Retrieval Prediction

Pre-retrieval prediction methods are used by retrieval systems to predict the quality of a ranked list of results retrieved in response to a query *without* actually retrieving the result list. Instead of considering the content of the result list, the methods rely on collection statistics and external resources such as semantic dictionaries to derive a prediction. The first research theme **RT1** revolves around these methods and considers the following research questions: On what heuristics are the prediction algorithms based? Can the algorithms be categorized in a meaningful way? How similar are different approaches with respect to their behavior to each other? How sensitive are the algorithms to a change in the retrieval approach? What gain can be achieved by combining different approaches?

We conclude in Chapter 2 that prediction methods are distinguished, in the literature, in four different classes according to the heuristics they exploit to predict the

effectiveness of a query. As such, *specificity* based prediction methods relate more specific query terms to a better performance, while *ambiguity* based predictors rely on the query terms' level of ambiguity to determine the performance. Ambiguous query terms are predicted to lead to a poor retrieval effectiveness while unambiguous query terms are viewed as evidence for a high retrieval effectiveness. A number of prediction methods also rely on the degree of *relatedness* between query terms to infer the query's performance: related query terms are predicted to lead to a better search result than unrelated query terms. Finally, the *ranking sensitivity* based prediction methods attempt to infer how difficult it will be for a retrieval approach to rank the documents that contain the query terms.

We performed an analytical and empirical evaluation of the prediction methods within each class and showed substantial similarities between them. When evaluating the prediction methods according to their ability to predict the effectiveness of queries on three different corpora we found their accuracy to be dependent on the retrieval approach, the query set and the corpus under investigation. We also showed that the dependency on the retrieval approach is very pronounced, not only when considering diverse retrieval approaches, but also when considering the different parameter settings of a single retrieval approach.

Overall, our results have indicated that when comparing predictor performances, a single retrieval setting can be misleading, and when possible, a variety of retrieval methods should be evaluated before conclusive observations are drawn about the merits of individual predictors. The general lack of predictor robustness as evinced from our work also brings into question the merits of pre-retrieval predictors; if they are unstable and often result in poor prediction accuracy, then the advantage of being low cost in terms of processing time is lost.

Finally the potential gain in accuracy when combining prediction methods has been explored. Specifically, we investigated the utility of penalized regression as a principled approach to combine predictors. The evaluation showed potential, for two of our three corpora the penalized regression methods led to improvements over the best single individual predictor.

### 6.1.2 Post-Retrieval Prediction

Approaches predicting the effectiveness of a query's result list by indeed considering the result list are employed after the initial retrieval stage and are thus called post-retrieval. The questions posed as part of the second research theme **RT2** were set around the post-retrieval predictor Clarity Score [45] and were as follows: How sensitive is this post-retrieval predictor to the retrieval algorithm? How does the algorithm's performance change over different test collections? Is it possible to improve upon the prediction accuracy of existing approaches?

In Chapter 3 we were able to show on two concrete predictor examples, one of which was Clarity Score, that post-retrieval prediction methods are as sensitive to the parameter settings of the retrieval approach as pre-retrieval predictors. The same observation holds for the performance of Clarity Score on different test corpora; the prediction accuracy varies widely depending on the corpus and the partic-

ular query set under investigation. We proposed two adaptations to Clarity Score: (i) setting the number of feedback documents used in the estimation of the query language model individually for each query to the number of documents that contain all query terms, and, (ii) ignoring high-frequency terms in the KL divergence calculation. These adaptations were thoroughly tested on three TREC test collections. With the exception of one set of queries, one or more of the proposed variations always outperformed the Clarity Score baseline, often by a large margin.

The main conclusion we draw from the investigations of Chapter 3 is that *Adapted Clarity* is a highly competitive post-retrieval approach which, on average across all evaluated corpora, outperforms all other tested pre- as well as post-retrieval predictors.

### 6.1.3 Contrasting Evaluation and Application

The third research theme dealt with the relationship of the current evaluation methodology for query performance prediction and the change in retrieval effectiveness of adaptive systems that employ a predictor for selective query expansion or meta-search. In selective query expansion a predictor is expected to predict when the application of automatic query expansion will lead to a higher quality result list. When applied in meta-search, for each query there exists a choice of result lists and a predictor is expected to identify the one of highest quality.

In particular the posed questions of **RT3** were: What is the relationship between the correlation coefficient as an evaluation measure for query effectiveness estimation and the effect of such a method on retrieval effectiveness? At what levels of correlation can we be reasonably sure that a query performance prediction method will be useful in an operational setting?

In Chapter 4 we provide first answers to these questions. We chose these two operational settings as they are most often mentioned as potential applications of query effectiveness prediction. Our experiments have shown that the level of Kendall's  $\tau$  required to be confident that a prediction method is viable in practice is dependent on the particular operational setting it is employed in. In the case of selective query expansion, a value of  $\tau \geq 0.4$  has been found to be the minimum level of correlation that should be attained provided perfect knowledge of the behavior of the employed automatic query expansion mechanism is available. A second experimental inquiry evaluated the effect of overly optimistic assumptions such as that query expansion aids all queries with initially high effectiveness. Under these circumstances predictors need to achieve a correlation of  $\tau \geq 0.75$  for them to be viable.

In the meta-search setting, the level of correlation required to reliably improve the retrieval effectiveness of a meta-search system is shown to be dependent on the performance differences of the participating systems as well as on the number of systems employed. Notably, when the effectiveness of all systems is similar, prediction methods achieving low levels of correlation are already sufficient. However, when the differences in system performance are large and we are interested in statistically significant improvements, the level of correlation necessary varies between

$\tau = 0.5$  ( $m = 150$ ) and  $\tau = 0.7$  ( $m = 50$ ) depending on the number  $m$  of queries participating in the experiment.

Based on the knowledge we gained in Chapter 2 and Chapter 3 we can convincingly state our main conclusion as follows: current query effectiveness prediction methods are not sufficiently accurate to lead to consistent and significant improvements when applied to meta-search and selective query expansion.

### 6.1.4 System Effectiveness Prediction

In Chapter 5 we turned to estimating the ranking of retrieval systems as set by the fourth research theme **RT4**. The questions posed were: Is the performance of system ranking estimation approaches as reported in previous studies comparable with their performance for more recent and diverse data sets? What factors influence the accuracy of system ranking estimation? Can the accuracy be improved when selecting a subset of topics to rank retrieval systems?

In order to answer these questions, we have investigated a wide range of data sets covering a variety of retrieval tasks and a variety of test collections. We found that in contrast to earlier studies which were mostly conducted on the same small number of data sets, there are indeed differences in the ability to rank retrieval systems depending on the data set. The issue that has long prevented this line of evaluation to be used in practice has been shown to be the mis-ranking of the best systems. In the extreme case, the most effective systems are estimated to be among the worst performing ones. In our experiments however, we have discovered this not to be an inherent problem of system ranking estimation approaches. The extent of the mis-ranking problem was shown to be data set dependent and, more specifically, dependent on the amount of human intervention in the best system of a data set. We conclude that in cases where the best system is (largely) automatic, the best system can often be identified with a high degree of accuracy.

The evaluation of retrieval systems has always been performed based on some set of topics. To answer the final question on accuracy improvement we first investigated the variability between topics, that is we evaluated how well the systems can be ranked for each individual topic. The result of this investigation motivated the follow up question on whether we can improve the ability of estimating a ranking of systems when relying on a subset of topics. In a motivational study we have shown that selecting topic subsets from the full set of topics can lead to a significantly higher accuracy.

The most important conclusion to have emerged from the work in Chapter 5 is that automatic system ranking estimation methods are *not* a lost cause. They are in fact capable of high quality estimations in contrast to previous findings.

## 6.2 Future Work

A number of future research avenues have become evident in the course of this work. One particular direction is the exploration of alternatives to the currently

employed correlation coefficients, namely Kendall's  $\tau$  and the linear correlation coefficient  $r$ . Blest [21] proposes a rank correlation coefficient that weights errors at the top end of the ranking more than errors at the bottom of the ranking, which stands in contrast to Kendall's  $\tau$  where all errors are weighted equally. In particular in the context of system ranking estimation, where we are often most interested in identifying the best performing systems correctly such an evaluation measure can be useful. In the context of Information Retrieval, Yilmaz et al. [166] propose a rank correlation coefficient based on average precision that penalizes errors at the top of the ranking to a higher degree. Both of the above may be considered as alternative evaluation measures in future work.

Most experiments in the realm of query performance prediction, including the experiments reported in this thesis, have been performed on informational queries. There are also other types of queries, such as navigational queries or transactional queries [23]. How the current approaches for informational queries can be translated to those query types largely remains an open question.

With the introduction of the Million Query track [3] to TREC, a much larger number of topics (10000 topics to be exact) has recently become available than the standard topic set size of between 50 and 250 topics for earlier test collections. Relevance judgments for such a large number of topics cannot be derived in the same manner as for a small set of topics though and, therefore, instead of average precision, new effectiveness measures had to be introduced such as statistical AP [8]. This development naturally leads to two further research questions; first to evaluate the performance of query performance prediction methods for such topic set sizes, and, second, to investigate if the novel evaluation measures can also be predicted reasonably well.

One future work prospect of Chapter 2 is the evaluation of the robustness of the prediction methods with respect to TREC runs. Since the predictor performances vary widely, it would be beneficial to analyze for which kind of retrieval approaches the different prediction methods perform well and for which they fail. Such an analysis would require an extensive review of all TREC runs and the methods they employ.

Future work related to Chapter 3 could focus on setting the feedback document parameter more effectively, specifically by taking into account the dependency between the query terms. Furthermore, the question of how best to set the  $N$  parameter automatically arises. An alternative line of investigation in particular for the Web corpus WT10g would be to preprocess the documents by filtering out the non-topical content such as navigational information, page decoration, etcetera. Such an approach has been shown to improve the effectiveness of pseudo-relevance feedback of the WT10g data set [168], and might also be beneficial for query effectiveness prediction.

A central assumption of Clarity Score is that for an unambiguous query, the top retrieved documents are more focused than the corpus. Although this is a valid assumption if each document contains exactly one topic, often documents covering multiple topics occur frequently in a collection and unnecessary noise is added to the query language model. Therefore, a future investigation could be the segmen-

tation of each document according to subtopics in order to alleviate these effects. The TextTiling [74] or C99 [35] segmentation algorithms could be employed for instance and those segmented passages may then be used where query terms occur in the creation of the query language model rather than relying on the entire document.

The study of Chapter 4, which investigated the contrast between the evaluation of query performance prediction and the application of prediction methods in practice, also offers diverse lines of follow-up research. In this work, we restricted ourselves to an analysis of Kendall's  $\tau$ , however, a future effort might perform a similar analysis of the linear correlation coefficient  $r$ . In contrast to the rank-based  $\tau$ ,  $r$  is based on raw scores, which adds another dimension to the study, namely the distribution of raw scores. A second measure which might also be investigated further is the *area between the MAP curves* [151], already briefly discussed in Section 2.3.2. Although it has not been widely used in the query performance prediction literature, we hypothesize that it is particularly useful for the operational setting of selective query expansion, as it emphasizes the worst performing topics.

Finally, the experiments on automatic system evaluation described in Chapter 5 could also be further explored. On the one hand, for topic subset selection to be beneficial, it is still necessary to develop an automatic method that identifies the most suitable subsets; here one could concentrate on identifying features that enable us to distinguish the topics that appear mostly in subsets which improve system ranking estimation, from those topics that appear mostly in poorly performing subsets. Another direction to consider is the adaptation of the *RS* approach by selectively boosting some documents in the pool of documents to sample from. This idea is motivated by the fact that in the case of easy topics, the very best systems will retrieve ranked lists of documents similar to average systems, while for more difficult topics the result lists will diverge. If we can identify the systems, that appear average on easy topics and unlike average systems on harder topics, we can boost the number of documents entered into the pool by them. This would require a comparison of document overlap across different topics, which is a deviation from current work where each topic is viewed in isolation.

Effectiveness predictors have great potential as adaptive systems that take the correct query-dependent actions are bound to outperform systems applying a one-size-fits-all approach. Although this potential is not yet fulfilled as shown in this thesis, current state-of-the-art methods are slowly beginning to reach the levels of accuracy required in practical settings, motivating future research in this direction.