A Case for Improved Evaluation of Query Difficulty Prediction

Falk Scholer Steven Garcia

School of Computer Science & IT
RMIT University, GPO Box 2476V
Melbourne, Australia, 3001
{falk.scholer,steven.garcia}@rmit.edu.au

ABSTRACT

Query difficulty prediction aims to identify, in advance, how well an information retrieval system will perform when faced with a particular search request. The current standard evaluation methodology involves calculating a correlation coefficient, to indicate how strongly the predicted query difficulty is related with an actual system performance measure, usually Average Precision. We run a series of experiments based on predictors that have been shown to perform well in the literature, comparing these across different TREC runs. Our results demonstrate that the current evaluation methodology is severely limited. Although it can be used to demonstrate the performance of a predictor for a single system, such performance is not consistent over a variety of retrieval systems. We conclude that published results in the query difficulty area are generally not comparable, and recommend that prediction be evaluated against a spectrum of underlying search systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Systems and Software**]: Performance Evaluation (Efficiency and Effectiveness)

General Terms

Experimentation, Measurement

1. EVALUATING PREDICTION

The prediction of query difficulty is a problem that has received considerable interest in the information retrieval community. The assumed benefit is that, if it is possible to predict in advance when a retrieval system is unlikely to return a useful set of answers for a query, then it would be possible to dynamically change the system's retrieval behavior (for example, through the selective application of query expansion, or by requesting additional information from the user).

In the information retrieval literature, a standard methodology for the evaluation of the effectiveness of a difficulty prediction technique has emerged. First, for each query in a set of test queries, a predicted difficulty score is calculated, based on one particular prediction approach (or *predictor*). For the same set of queries, a system performance metric (usually Average Precision) is calculated. Finally, a correlation coefficient is calculated between these two sets of numbers, to quantify the strength of the relationship between the predicted and actual system performance. The higher the correlation coefficient, the "better" the predictor is considered to be. In the literature, three correlation coefficients are commonly used—Pearson (linear) correlation, Spearman (rank) correlation, and Kendall's tau (also based on ranks)—although papers often report only a subset of these.

In addition to the value of the correlation coefficient itself, it is usual to report a p-value, representing the outcome of a statistical inference test of the null hypothesis that the correlation has occurred by chance [5]. With a sample size of 50 queries, relatively weak correlations with a coefficient of around 0.2 are usually enough to establish statistical significance at the 0.05 level; where larger query sets are evaluated, even lower correlation values are significant. Typically, if the null hypothesis can be rejected at a suitable level of significance, and if the strength of the correlation coefficient is comparable to or higher than a competing prediction method, then the current approach is reported to be a success [1, 3, 6].

2. EXPERIMENTS AND DISCUSSION

This evaluation process, while seemingly straightforward, actually requires the experimenter to make a variety of choices, and includes a range of strong assumptions, which are often not reflected in the interpretation of the experimental results. A particularly serious issue is that predictors are usually evaluated in comparison to only a single retrieval system. As we demonstrate, this issue limits the methodology to providing informative outcomes only about specific predictor-system combinations (whereas the results of such experiments are often treated as providing some sort of general information about the applicability of a prediction approach), and in particular makes it dangerous to draw conclusions from results between different implementations (something that is often done in practice).

We compare the performance of three different query difficulty prediction approaches: Query Clarity (QC) [1], Maximum IDF (MI) [4], and Maximum Variability (MV) [6]. These are representative of approaches that are used widely in the literature [2]. The first is a post-retrieval approach, and makes use of information from the set of documents that is returned in response to the initial query, while the lat-

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

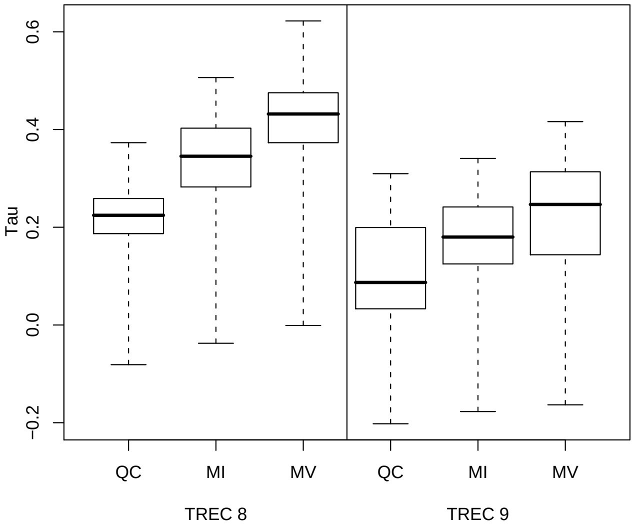

Figure 1: Distribution of Kendall's tau values for Query Clarity (QC), Maximum IDF (MI), and Maximum Variance (MV) on the TREC 8 and 9 collections.

ter are pre-retrieval approaches, based on term-distribution statistics.

To examine the effect that varying the underlying retrieval system has on the reported performance of a predictor, the three predictors are evaluated on each run submitted for two TREC *ad hoc* tasks: TREC 8 (topics 401–450, newswire data, 129 submitted runs), and TREC 9 Web track (topics 451–500, Web data, 105 submitted runs). Following the usual evaluation approach, we measure predictor performance by correlating the predicted scores against Average Precision (due to space limitations, we report results using Kendall's tau only).

The range of tau values obtained when using the three prediction approaches across all retrieval runs submitted for TREC 8 and TREC 9 are shown in Figure 1. The performance of each predictor varies dramatically depending on the underlying retrieval system (TREC run) that the prediction scores are being correlated with. For example, the performance of the *MI* predictor on TREC 9 data, as measured by correlation with AP, ranges from -0.12 to 0.44. In other words, the predictor varies from having a strong positive relationship with some systems, to having a negative relationship with others. Moreover, runs with tau values of less than around 0.2 are not statistically significant at the 0.05 level. For the TREC 9 data, for example, this means that for each predictor, around one half of the runs produce correlations that are not significant.

These results are striking: just because an approach works well with one chosen retrieval system does not imply that it will work well with others. The current methodology, where predictors are typically evaluated against only one retrieval system (a single TREC run), therefore does not allow general conclusions about the performance of a prediction technique to be drawn.

	QC & MI	QC & MV	MI & MV
TREC-8	0.3846	0.4129	0.1422
TREC-9	0.3198	0.2920	0.2690

Table 1: The likelihood with which the relative performance of a pair of predictors will change, for any two randomly chosen runs.

More worryingly, the framework may lead directly to contradictory results. The correlation scores are used in the current evaluation framework as direct measures of predictor performance, and are thus used to compare one predictor with another. However, depending on the choice of run with which the predictors are being evaluated, this relative ordering may change. For example, consider an attempt to compare the performance of the *MI* and *MV* predictors. When these predictors are correlated with one underlying retrieval system, *MI* may show a higher correlation with the system AP scores. However, when another retrieval system is used to calculate AP, it may be that *MV* leads to a stronger correlation. Table 1 shows the likelihood that the relative performance of two predictors would be reversed when the performance is calculated using any two randomly chosen runs. For TREC 8, contradictory conclusions would be reached from 14% to 41% of the time, while for TREC 9 the likelihood ranges from 27% to 32%.

3. CONCLUSIONS

The main drawback of the current evaluation approach for query difficulty prediction is that, instead of leading to conclusions about the general performance of a predictor, it is limited to providing information about performance in relation to a single retrieval system. As our analysis has shown, such performance can vary widely from system to system, and therefore any claims about the general performance of predictors beyond the context of the particular retrieval system that they were evaluated with should be viewed with caution. As a minimum, evaluations of predictor quality should show correlations with a number of representative underlying retrieval systems.

We continue to believe that query difficulty prediction is a worthwhile problem, but that a stronger evaluation framework is required for this area of research to yield meaningful conclusions. In future work, we plan to investigate prediction evaluation approaches that take a broader view of query difficulty.

4. REFERENCES

[1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. SIGIR*, pages 299–306, Tampere, Finland, 2002.

[2] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proc. CIKM*, pages 1419–1420, Napa Valley, CA, 2008.

[3] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31(7):585–594, 2006.

[4] F. Scholer, H. E. Williams, and A. Turpin. Query association surrogates for web search. *JASIST*, 55(7):637–650, 2004.

[5] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures.* CRC Press, 1997.

[6] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. ECIR*, pages 28–39, Glasgow, UK, 2008.