Predicting Query Performance for User-based Search Tasks

Ying Zhao *Falk Scholer*
School of Computer Science and IT
RMIT University
Melbourne, Australia
{*ying.zhao,falk.scholer*}*@rmit.edu.au*

Abstract *Query performance prediction aims to determine in advance whether a user's search request will return a useful answer set. The success of such prediction attempts are currently evaluated by calculating the correlation between the predicted performance and standard information retrieval metrics of system performance such as average precision. However, recent work suggests that there is little relationship between average precision and the performance of users when carrying out search tasks. Direct measures of user performance offer another way of evaluating the effectiveness of search systems; this is of particular importance in the framework of query prediction, since one of the goals of prediction is to warn users when search results are likely to be poor. We therefore investigate the relationship between current prediction techniques and user-based performance measures. Our preliminary results show that the performance of the predictors differs strongly when using system-based compared to user-based performance measures: predictors that are significantly correlated with one measurement are often not correlated with the other. In general, the predictors are more correlated with average precision rather than with user performance.*

Keywords Query performance prediction, information retrieval, user study

1 Introduction

Query performance prediction has a wide range of potential applications that aim to improve search performance, seeking to provide users with knowledge about whether a set of search results are likely to contain useful answers for their information needs [2, 8]. The assumed benefit of prediction stems from the fact that the behaviour of a retrieval system could be changed dynamically based on the expected success of the query. For example, when a user enters a search request that is likely to lead to poor answers, the search system might prompt the user to re-formulate their query, without the user first having to work their way through a poor result list.

Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007.

Current prediction techniques can be grouped into three categories: pre-retrieval prediction [4], post-retrieval prediction [11, 12], and learning prediction [9]. Despite differences in the prediction approaches, the evaluation of such techniques follows a common methodology: a set of topics is run over a chosen collection and, based on available relevance judgements, a per-topic performance metric is calculated. A prediction technique is then used to estimate the performance of each of the topics. The correlation between the predicted and "actual" performance values is then calculated, usually together with a statistical test to demonstrate that the correlation is significant. The higher the correlation, the better the predictor is deemed to be.

The "actual" performance metric that is to be predicted is a precision-based measurement; in all recent studies of query performance prediction the *average precision* (AP) of the search system [3, 4, 11, 12]. However, it has been shown that there is little relationship between AP scores and actual user performance on a variety of search tasks [7]. Since query performance prediction aims to support the *user* in resolving an information need, it is important to investigate whether current prediction techniques behave differently when considering user-based measures of search performance.

In this work, we report on initial experiments that examine prediction techniques using both the conventional evaluation metric (AP), and user-based performance. A total of 9 predictors are tested from different perspectives; the results demonstrate that using AP and user-based performance—different ways of measuring search system effectiveness—results in different evaluation outcomes for prediction techniques.

2 Background

Given a query $Q(t_1, \ldots, t_n)$ and a collection C, prediction methods compute a score to estimate the performance of a query. Pre-retrieval predictors are calculated based on information that is available at indexing time; there is no need to the search system to evaluate the full result set for a query, giving advantages in terms of simplicity and efficiency. Queries are distinguished by exploring term statistics and distributions in the col-

lection. Several pre-retrieval predictors have been proposed in recent years, and we investigate 9 state-of-art predictors in this preliminary study. Since this work focuses on methodological aspects of prediction, we only provide brief details of the predictors themselves.

Clarity evidence. He and Ounis proposed a variety of pre-retrieval predictors [4]; based on their experimental results, the simplified clarity score (SCS) and average inverse collection term frequency (AveICTF) showed the best performance. SCS is a variation of the classic clarity score, a post-retrieval predictor originally proposed by Cronen-Townsend et. al [3].

$$SCS = \sum_{t \in Q} \theta_q \cdot \log_2 \frac{\theta_q}{\theta_c}$$

$$AveICTF = \frac{|C|}{f_{c,t}}$$

where θ represents a language model (see He and Ounis [4] for estimation details); N is the number of terms in the collection; $|C|$ is the number of documents in the collection; and $f_{c,t}$ is the term frequency within the collection. We also consider the maximum inverse document frequency (MaxIDF) as a predictor [5]:

$$MaxIDF = \frac{N}{f_t}$$

where f_t is the number of documents that contain t.

Similarity evidence. This family of predictors computes a similarity score between a query vector and a collection vector [10]. The *SCQ* score combines evidence from the frequency with which terms occur in the collection, and the inverse document frequency. We also consider two variations—the normalised score (*NSCQ*), and the maximum SCQ score (*MaxSCQ*). The intuition behind *MaxSCQ* is that, since web search queries tend to be short, if at least one of the terms has a high score then the query as a whole can be expected to perform well. The three predictors are defined as:

$$SCQ = \sum_{t \in Q} (1 + \ln(f_{c,t})) \times \ln\left(1 + \frac{N}{f_t}\right)$$

$$NSCQ = \frac{SCQ}{|Q|_{t \in \mathcal{V}}}$$

$$MaxSCQ = argmax\left[\forall_{t \in Q}\ SCQ_t\right]$$

Variability evidence. Variability evidence is collected as the term distribution over the entire collection [10]. The standard deviation is a statistical measure of dispersion, which reflects how widely spread the values in a data set are around the mean. In the context of prediction, intuitively if the standard deviation of the distribution of term weights over the entire collection is low, then the retrieval system will be less able to distinguish between highly relevant and less relevant documents, and the query is therefore likely to be more difficult to evaluate. Again, we also consider the normalised and maximum versions of this measure:

$$\sigma_1 = \sum_{t_1}^{t_n} \sqrt{\frac{1}{f_t} \sum_{d \in \mathcal{D}_t} \left(w_{d,t} \overline{w}_t\right)^2}$$

$$\sigma_2 = \frac{\sigma_1}{|Q|_{t \in \mathcal{V}}}$$

$$\sigma_3 = argmax\left[\forall_{t \in Q}\ \sigma_{1,t}\right]$$

where $\overline{w}_t$ is the mean in-document term occurrence, $\frac{\sum_{d \in \mathcal{D}_t} w_{d,t}}{N}$.

3 Experimental Methodology

We investigate the difference in the performance of predictors when evaluated based on a standard IR metric, average precision (AP), and when using user-based measures. For our experiments we use the WT10g collection—a 10Gb crawl of the web in 1997 [1]. This collection was used in the TREC 9 and 10 Web tracks, and has a corresponding set of TREC topics and relevance judgements, with which the AP of retrieval systems can be calculated. For user-based measures, we use data collected in a user study by Turpin and Scholer [7]. 30 subjects conducted searches on the WT10g collection using 47 queries, and 5 different search systems, where each system returned answer lists at a pre-determined AP level.

A precision-oriented user-based measure of performance is the length of time that it takes to find the first relevant document for an information need. Analysis of the user data indicated statistically significant user effects (that is, some users take consistently longer than others to find the first answer), as well as statistically significant topic effects (across all search systems and users, it takes significantly longer to find a relevant answer for some topics than for others) [7]. As a user-centric measure of query difficulty, we use the average time required to find a relevant document for a topic. Since the raw time data does not appear to be normally distributed, we use the median time required to find an answer to measure the system effectiveness.

In the query performance prediction literature, three different correlation coefficients are widely used: the Pearson product-moment correlation; Spearman's rank order correlation; and, Kendall's tau. Correlation coefficients vary in the range $[+1, -1]$; a value of zero indicates that there is no relationship between the two groups of data. For each correlation, a statistical test can be performed to test whether the relationship is significant at a specified level of confidence (in this paper we use a standard significance level of 0.05). For a comprehensive treatment of the properties of the different correlation coefficients the reader is referred to Sheskin[6].

We note that there is no consensus about which correlation coefficient is the most appropriate for query performance prediction; individual papers often report

Table 1: *Pearson, Kendall, and Spearman correlation coefficients between user-based difficulty measure (median time to find a relevant document) and pre-retrieval predictors. Bold entries indicate that the correlation is statistically significant at the 0.05 level.*

	1	2	3	4	5	6	7	8	9
Predictor	SCS	AveICTF	MaxIdf	SCQ	NSCQ	MaxSCQ	σ_1	σ_2	σ_3
Pearson	0.107	-0.122	**0.390**	0.257	-0.127	0.208	**0.355**	0.142	**0.343**
p-value	*0.480*	*0.419*	*0.007*	*0.084*	*0.399*	*0.165*	*0.015*	*0.347*	*0.019*
Kendall	0.138	-0.098	0.197	0.163	-0.080	0.143	**0.208**	0.076	0.166
p-value	*0.179*	*0.346*	*0.053*	*0.112*	*0.428*	*0.161*	*0.042*	*0.462*	*0.103*
Spearman	0.204	-0.147	0.288	0.234	-0.128	0.212	**0.313**	0.112	0.236
p-value	*0.173*	*0.330*	*0.052*	*0.117*	*0.396*	*0.157*	*0.035*	*0.454*	*0.115*

only one or two of the available variants. However, the choice of correlation coefficient can lead to different conclusions about the performance of a predictor. Therefore in this study, we apply all three correlation methods to examine whether different test methods lead to different results with the same data sets.

4 Results and Discussion

We first analyse the correlation between the pre-retrieval predictors of query difficulty and our user-based measure of topic difficulty, the median time to find the first relevant document. The results are shown in Table 1. The σ_1 predictor is the most strongly correlated, and is the only predictor for which the correlation is statistically significant across all three correlation co-efficients. The correlations of the $MaxIDF$ and σ_3 predictors are significant only with Pearson correlation, and the other predictors show no significant relationship at all with the user-based measure of topic difficulty.

All of the selected predictors have been reported to be significantly correlated with a system-based performance measure, AP. However, these results are based on different collections or topics [4, 10]. We therefore conducted a second experiment, running the 47 topics for which we have user data as standard queries (using TREC title fields only), and calculating the AP of system performance on each. In this experiment, we used the Indri search engine with Dirichlet smoothing, where μ is set to 1000[1].

We compare the results of using different system performance metrics (AP and median search time) in evaluations of query prediction techniques. The results are presented in Figure 1 (a) to (c) for the three correlation coefficients. The numbers from 1 to 9 on the x-axis correspond to the 9 predictors (the ordering is the same as in Table 1). The y-axis shows the correlation coefficients. For each predictor, a pair of bars are shown: the left bar indicates the correlation with AP, and the right shows the correlation with the median time to find the first relevant document. Statistically significant correlations are shown as in dark gray, while non-significant results are shown as light gray.

[1]Indri is available from `www.lemur.org`.

There is a lack of consistency in the correlation results using the different difficulty measures: across all 3 correlation types, there are many cases (for example $MaxSCQ$ and σ_2) where a predictor is significantly correlated with AP but not with median time. The main exceptions are some poorly performing predictors such as SCS and SCQ that don't have a relationship with either difficulty measure. The only predictor that has a significant relationship with both measures is σ_1.

We note that not only the choice of difficulty measure, but also the choice of correlation technique, can affect the conclusions about predictor performance. For example, the Pearson correlation shows that $MaxIDF$ is significantly correlated with the median time measure; however, the other two correlations suggest that the relationship is not significant. Despite the inconsistencies observed, in general the selected predictors are more correlated with the average precision rather than user performance.

In a third experiment, we examine the direct relationship between the two measures of difficulty that we have used. The results for all three correlation coefficients are shown in Figure 1 (d). Although the Pearson correlation indicates a statistically significant relationship, the two rank-based correlation techniques do not identify a significant effect. The cause of this inconsistency may be an underlying assumption of the Pearson correlation, namely that the relationship is linear. There is no *a priori* reason to believe that a linear relationship should hold between the average precision of a retrieval system and the time taken to find the first relevant document. The results therefore do not provide evidence of a significant relationship between the user- and system-based measures of topic difficulty.

5 Conclusions

In this paper, we have conducted a preliminary investigation of using user-based measures of query difficulty for evaluating the effectiveness of query performance predictors. Experiments with 9 state-of-the-art pre-retrieval predictors showed that there is a lack of consistency in correlation; in general, previously proposed predictors tend to be related with only one or none of the difficulty measures. The only exception is

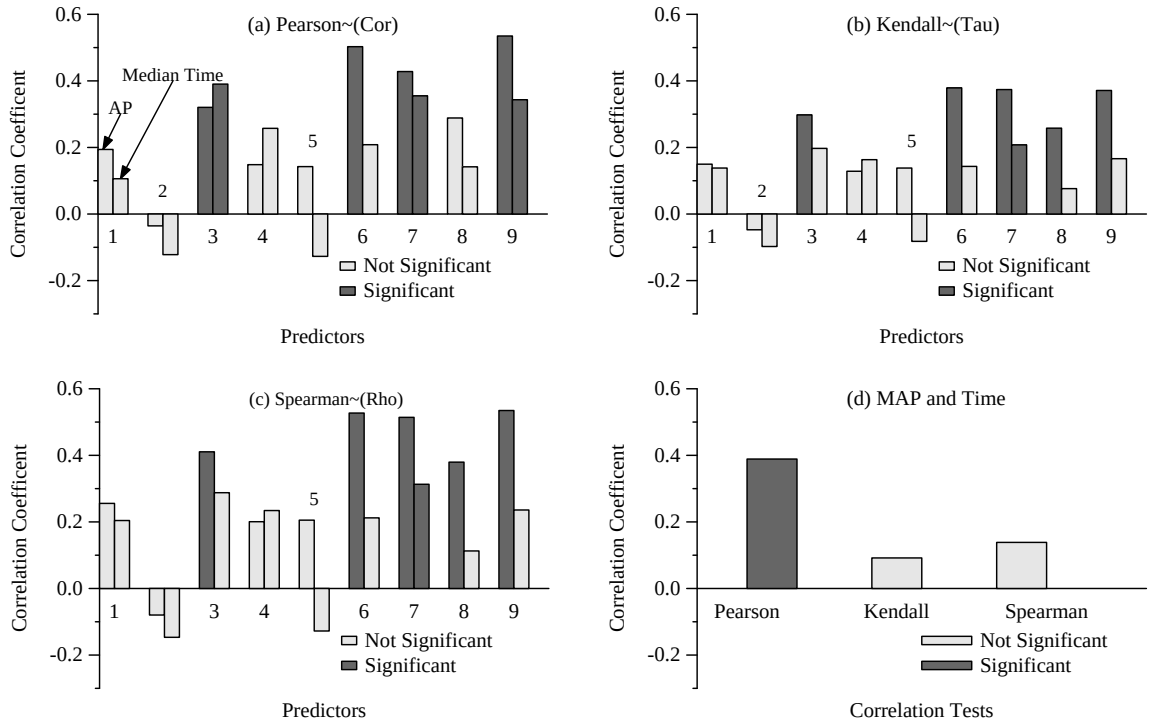

Figure 1: *The evaluation of prediction effectiveness using both average precision and user performance data. Three correlation tests are applied: Pearson shown as figure (a), Kendall as shown in figure (b), and Spearman as shown in figure (c). The relationship between two system performance metrics is shown in figure (d). In all figures, bars in dark gray indicate the results statistically significant at 0.05 confidence level.*

the σ_1 predictor—based on the variability of the distribution of query terms across documents—which had a significant relationship with both measures.

These findings have implications for the methodology that is currently used to evaluate the effectiveness of query performance predictors—it is not clear that simply showing a correlation with AP, as has been standard in the literature, is a suitable reflection of the actual difficulty that *users* face when searching on different topics.

In future work we plan to expand our investigation to include post-retrieval predictors, and to further examine the query performance prediction methodology. We also plan to investigate the use of the different correlation coefficients more thoroughly: since these sometimes give conflicting results, it needs to be established which measures are the most appropriate in the context of query performance prediction.

References

[1] P. Bailey, N. Craswell and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, Volume 39, Number 6, pages 853–871, 2003.

[2] D. Carmel, E. Yom-Tov and I. Soboroff. SIGIR workshop report: predicting query difficulty—methods and applications. *SIGIR Forum*, Volume 39, Number 2, pages 25–28, 2005.

[3] S. Cronen-Townsend, Y. Zhou and W. B. Croft. Predicting query performance. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[4] B. He and I. Ounis. Query performance prediction. *Information System*, Volume 31, Number 7, pages 585–594, 2006.

[5] F. Scholer, H. E. Williams and A. Turpin. Query association surrogates for web search. *American Society for Information Science and Technology*, Volume 55, Number 7, pages 637–650, 2004.

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

[7] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–18, Seattle, USA, 2006.

[8] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, 2006. National Institute of Standards and Technology Special Publication 500-266.

[9] E. Yom-Tov, S. Fine, D. Carmel and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 512–519, Salvador, Brazil, 2005.

[10] Y. Zhao, F. Scholer and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In submission.

[11] Y. Zhou and W. B. Croft. Ranking robustness: a novel framework to predict query performance. In *Proceedings of the 15th ACM CIKM International Conference on Information Knowledge Management*, pages 567–574, Arlington, Virginia, USA, 2006.

[12] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 543–550, Amsterdam, The Netherlands, 2007.