Query Performance Prediction By Considering Score Magnitude and Variance Together

Yongquan Tao[1] Shengli Wu[1,2]
[1]School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China
[2]School of Computing and Mathematics, Ulster University, Newtownabbey, UK
taoyongquan77@126.com,swu@ujs.edu.cn

ABSTRACT

Query Performance prediction aims to evaluate the effectiveness of the results returned by a search system in response to a query without any relevance information. In this paper, we propose a method that considers both magnitude and variance of scores of the ranked list of results to measure the performance of a query. Using six different TREC test sets, we compare our predictor with three of the state-of-the-art techniques. The experimental results show that our method is very competitive. Pairwise comparisons with each of the three other methods show that our predictor performs better in more data sets.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

query difficulty; query performance prediction; score distribution

1. INTRODUCTION

Query difficulty prediction aims to predict whether a query will return a high quality result list ("easy" queries), or low quality result list ("hard" queries), when no relevance information is given by a human operator. Query difficulty prediction is also referred to as query performance prediction. It has many potential applications in a variety of IR tasks such as improving retrieval consistency, query refinement, and distributed IR. This is why the problem has received considerable attention in the IR community in recent years.

Accurate performance predictions can help a user decide if the results are acceptable. If more relevant results are needed, then the user may decide to reformulate the query so as to obtain some different results from the same search engine as before or use other search services available.

Query performance prediction can be roughly categorized into two types: pre-retrieval prediction and post-retrieval prediction. Pre-retrieval methods evaluate the query before the search takes place, thus they must rely on the statistics of the query terms in the collection [8]. The advantage of such methods is that they can be computed quickly, using available statistics of the query terms gathered at indexing time. However, a disadvantage of such predictors is that they do not take into account the specific retrieval algorithms, so the predictions may not be as accurate as the post-retrieval prediction methods [3].

Post-retrieval prediction methods are usually more complex and expensive as the search results need to be analyzed after retrieval. Post-retrieval prediction algorithms can be further divided into clarity score based methods [10], ranking robustness based methods [14], and score analysis based methods [12, 13].

Prior research on score analysis demonstrates that magnitude and variance of scores are two factors that are correlated with query performance. The score-based performance prediction methods proposed previously consider either magnitude or variance of scores [13, 2, 9] but not both. In this paper, we propose a method that takes both magnitude and variance of scores into consideration at the same time. Experiments with 6 groups of TREC data are very promising..

The rest of this paper is organized as follows: related prior work is discussed in section 2. Section 3 describes our method for creating an estimator for query performance in detail. Section 4 presents the experiments we conducted on TREC data. Section 5 concludes the paper.

2. PRIOR WORK

In this section, we review some prior work on post-retrieval query prediction which category our proposed method falls into.

Cronen-Townsend et al.[10] proposed a method of computing the relative entropy between the models of the query and the collection. Afterwards, a few more clarity-based predictors have been proposed by other researchers [5, 6, 7].

Zhou et al.[14] built a novel framework called ranking robustness to predict query performance. Robustness based approaches evaluate how robust the results are to perturbations in the query, the result list and the retrieval method.

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661906.

Related research by others may be found, for example, in [1, 4, 11].

A main branch of the post-retrieval prediction methods is score analysis. Zhou et al. proposed a predictor, Weighted Information Gain (WIG) [13], which measures the divergence between the mean retrieval score of some top-ranked documents and that of a typical document in the entire corpus. Shtok et al. proposed another predictor called Normalized Query Commitment (NQC) [2] to estimate query drift in the list of top-ranked and/or bottom-ranked documents. More recently, Pérez-Iglesias et al. also focused on the variance aspect of scores. Some experiments are conducted by using standard deviation and some of its variants to capture the differences between "hard" and "easy" queries. [9].

In this paper, we investigate the problem of query performance prediction by analyzing score distribution. Those methods based on score analysis focused on either the magnitude [13] or the variance [2, 9] of the scores, yet none are able to utilize both. Therefore, we propose a query performance prediction method that takes both factors into consideration at the same time. As we can see later, such a combination is not trivial. Experiments with TREC data are conducted to evaluate the effectiveness of our method.

3. METHODOLOGY

Let us begin by setting out the notation. Let q, $\mathcal{D}$, and $\mathcal{M}$ denote a query, a corpus, and a retrieval method, respectively. We use $L(q, \mathcal{M})$ and $\mathcal{D}_q^{[k]}$ to denote the result list returned in response to query q by $\mathcal{M}$ over $\mathcal{D}$ and the top-k documents ranked highly in the result list $L(q, \mathcal{M})$, respectively. k is a free parameter, set to an arbitrary natural number prior to the search. Our goal is to establish a predictor for evaluating the quality of the ranking list returned by $\mathcal{M}$ over $\mathcal{D}$ for a given query q without relevance judgment information.

Previous work on query performance prediction observes that there is a certain relationship between score distribution and query performance [2, 9, 13]. In particular, two factors, namely the magnitude and deviation of scores may be used to predict query performance.

WIG [13] uses Equation 1

$$WIG(q, M) = \frac{1}{k} \sum_{d \in \mathcal{D}_q^{[k]}} \frac{1}{\sqrt{|q|}} (Score(d) - Score(\mathcal{D})) \quad (1)$$

to calculate scores for a given query q. [1] Here $Score(d)$ is the score that document d is awarded by $\mathcal{M}$, $Score(\mathcal{D})$ is the score that an average document in $\mathcal{D}$ would be given by $\mathcal{M}$, and $|q|$ is the number of terms in q. In Figure 1, we can see that WIG mainly considers the magnitude of scores that those retrieved documents obtain. $\sqrt{|q|}$ serves as an scale factor to make WIG scores comparable over different queries.

NQC uses Equation 2

$$NQC(q, \mathcal{M}) = \frac{1}{Score(\mathcal{D})} \sqrt{\frac{1}{k} \sum_{d \in \mathcal{D}_q^{[k]}} (Score(d) - \hat{\mu})^2} \quad (2)$$

[1]This is a simplified version of WIG, which only uses score information of the results. According to [12], it is a very effective method.

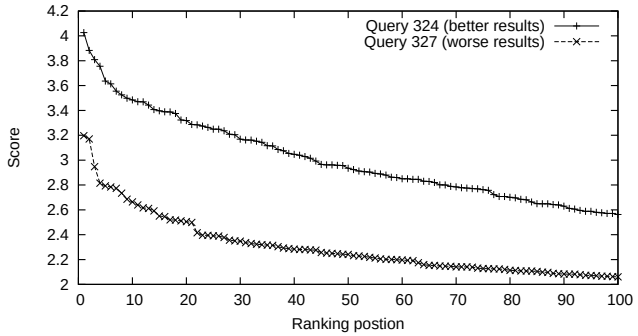

Figure 1: Topic 324 "Argentine/British Relations", average precision is 0.6670; topic 327 "Modern Slavery", average precision is 0.1773. Data is taken from $pircRB04t3$, a run submitted to the Robust Track in 2004.

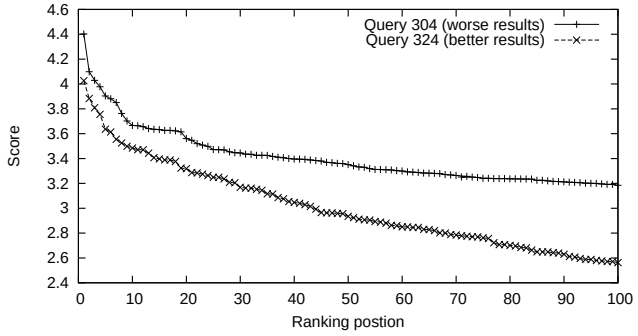

Figure 2: Topic 324 "Argentine/British Relations", average precision is 0.6670; topic 304 "Endangered Species (Mammals)", average precision is 0.1049. Data is taken from $pircRB04t3$, a run submitted to the Robust Track in 2004.

to calculate scores for a given query q. Here $\hat{\mu}$ is the average of the scores of all k results in $\mathcal{D}_q^{[k]}$. NQC [2] can be regarded as a variation of standard deviation (referred to as SD), which is investigated in [9].

Let us consider two examples to illustrate why WIG and NQC work in some situations, but fail in others. All the data is taken from $pircRB04t3$, which is the best (measured by MAP) among all those submitted to TREC Robust 2004. Figure 1 shows the results for queries 324 and 327. The results for query 324 obtain higher scores than that for query 327, and the variance of both score distributions are similar. MAP(324)>MAP(327). These conditions are ideal for performance prediction methods such as WIG. On the other hand, methods such as SD and NQC may not work well since the curves for both queries have very similar shape. As a matter of fact, we have

$$NQC(324, pircRB04t3) = 0.202$$

and

$$NQC(327, pircRB04t3) = 0.210$$

so NQC would make a wrong decision about performance comparison of these two queries.

The second example is to compare 324 "Argentine/British Relations"and query 304 "Endangered Species (Mammals)".

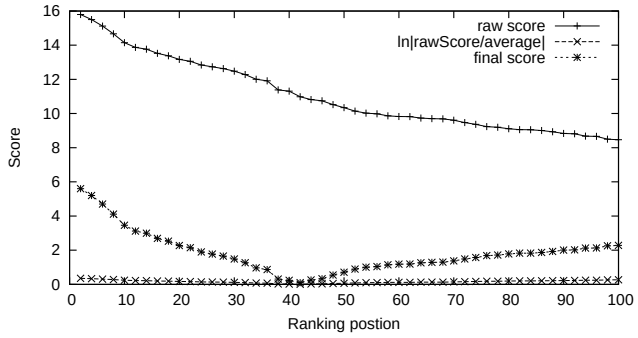

Figure 3: Topic 561. Data is taken from $pircRB04t3$, a run submitted to the Robust Track in 2004.

MAP(324) >MAP(304). Figure 2 shows the score curves for both results. The results for query 304 obtain higher scores than that for query 327, but the latter has a greater variance than the former. These conditions are ideal for performance prediction methods such as NQC and SD, but they are not good for methods such as WIG.

From the above two examples we can see that considering either score magnitude or deviation may work in some situations but not the others. It would be an advantage if we can consider both of them together. We use the following Equation 3

$$SMV(q, \mathcal{M}) = \frac{\frac{1}{k}\sum_{d \in \mathcal{D}_q^{[k]}} (Score(d)|\ln\frac{Score(d)}{\hat{\mu}}|)}{Score(\mathcal{D})} \qquad (3)$$

to calculate scores for a given query q. Here SMV stands for our method, which considers both Score Magnitude and Variance. $\hat{\mu}$ is the average of the scores of all k results in $\mathcal{D}_q^{[k]}$. Inside the summation of Equation 3, there are two components. One is $Score(d)$ and the other is $|\ln\frac{Score(d)}{\hat{\mu}}|$. The former is used to represent score magnitude and the latter is used to represent a form of score variance. We combine the two by multiplication.

In order to understand the contribution that results at different ranks can make, let us consider an example, the results for query 561 from $thutd5$. At each rank, the value of raw score $Score(d)$, the value of $w(d) = |\ln\frac{Score(d)}{\hat{\mu}}|$, and the final score (product of $w(d) * Score(d)$) are shown in Figure 3. We can see that each result makes some contribution to the final score. Results at the very top and bottom make more contribution than those in the middle. This is reasonable because results at both ends are more informative than those in the middle. Let us recall the two aforementioned examples. WIG works in the first example, but not the second; NQC works in the second example, but not the first. SMV works in both examples.

In its current form, Equation 3 works with positive scores but not negative scores. If there are negative scores, then Equation 3 can be modified to support that. Let us define low_s to be the lowest score from the results for a group of queries. $Score(d)$ can be replaced by $Score(d) - low\_s$, thus all negative scores transform to positive scores and Equation 3 can be used without any problems.

[2]Topic 672 is removed because of no relevant results identified.

Table 1: Summary of test collections

Collection	Size	TREC Task	Topics
Disks 2&3	567,529	Ad hoc 4	50
Disks 2&4	524,929	Ad hoc 5	50
Disks 4&5-CR	528,155	Robust 2004	249 [2]
WT10G	1.69 m	Web 2001	50
GOV	1.25 m	Web 2002	50
GOV2	25.2 m	Terabyte 2004	50

Table 2: $Pearson's$ correlation coefficients for correlation with actual retrieval performance. Bold cases mean the most accurate prediction per collection.

TREC TASK	SD	WIG	NQC	SMV
TREC 4	0.418	**0.505**	0.414	0.442
TREC 5	0.296	0.474	0.573	**0.586**
ROBUST 2004	**0.646**	0.575	0.597	0.591
WebTrack 2001	0.359	0.312	0.329	**0.397**
WebTrack 2002	0.330	0.295	0.444	**0.452**
Terabyte 2004	0.373	0.307	0.493	**0.521**

4. EVALUATION

In this section, we evaluate SMV. Experiments are conducted on 6 different TREC collections. Table 1 summarizes the information of these test collections. They are used in different tasks including ad hoc, web, terabyte, and robust. Their sizes vary from 0.5 million (disks 2&3, disks2&4, disks 4&5-CR) to 25 million (GOV2). We compare the prediction quality of SMV with that of the three state-of-the-art predictors: Standard Deviation (SD) [9],Weighted Information Gain (WIG) [13] and NQC [2].

As described in previous section, we need to set a value for the parameter k in all the predictors. As recommended in [12], k is set to 5 for WIG. For the three other methods NQC, SD and SMV, k is set to the same value. It is set to 1000 for the GOV2 collection and 100 for all other collections, as in [2]. Special treatment is given to GOV2 because in the TREC 2004 Terabyte task, 10000 results were retrieved for each query, whereas only 1000 results were retrieved for all other cases. We use the average score of all the retrieved results (1000 or 10000) to estimate $Score(D)$.

For each collection, we select the best run (measured by MAP) that is submitted to TREC to carry out the experiment. They are $CnQst2$ (TREC 4), $ETHme1$ (TREC 5), $fub01be2$ (TREC 2001), $thutd5$ (TREC 2002), $pircRB04t3$ (TREC 2004 Robust), $uogTBQEL$ (TREC 2004 Terabyte).

The prediction quality of a method is evaluated by measuring both $Pearson's$ and $Kendall's - \tau$ correlation between the ranking of queries by their actual performance (measured by MAP) and the ranking of queries by a performance predictor. In statistics, $Pearson's$ correlation is a measure of the linear correlation between two variables. Its range is [-1,1], where 1 presents total positive correlation, 0 no correlation and -1 total negative correlation. $Kendall's - \tau$ coefficient is used to measure the association between two measured quantities. Its range is also [-1,1], where 1 denotes that the two rankings are the same, and -1 denotes that one ranking is the reverse of the other.

The experimental results are shown in Tables 2 and 3 for $Pearson's$ correlation and $Kendall's - \tau$ rank coefficient

Table 3: $Kendall's-\tau$ rank coefficients for correlation with actual retrieval performance. Bold cases mean the most accurate prediction per collection.

TREC TASK	SD	WIG	NQC	SMV
TREC 4	0.333	**0.352**	0.269	0.299
TREC 5	0.229	0.350	**0.427**	0.425
ROBUST 2004	**0.444**	0.404	0.394	0.396
WebTrack 2001	0.279	0.230	0.273	**0.301**
WebTrack 2002	0.139	0.158	0.177	**0.194**
Terabyte 2004	0.243	0.185	0.340	**0.364**

Table 4: Average of $Pearson's$ correlation (P) and $Kendall's-\tau$ rank coefficients (K) for correlation with actual retrieval performance. 6 collections are divided into two types: clean (including Disks 2&3, Disks 2&4 and Disks 4&5-CR) and noisy (including WT10g, GOV and GOV2).

Collections	SD	WIG	NQC	SMV
Clean	0.453(P)	0.518(P)	0.528(P)	0.540(P)
	0.336(K)	0.369(K)	0.363(K)	0.373(K)
Noisy	0.354(P)	0.305(P)	0.422(P)	0.457(P)
	0.220(K)	0.191(K)	0.263(K)	0.286(K)

respectively. Generally speaking, SMV predicts query performance better in more collections when either of the two measures is used. More specifically and compared in pairwise fashion with any of the three other methods, SMV outperforms each of the three methods in 5 out of 6 collections with respect to $Pearson's$ correlation; the figures are 4 (SD), 4 (WIG), and 5 (NQC) out of 6 if considering $Kendall's-\tau$ rank coefficient. Apart from the best run in each task, we also randomly select and evaluate a few more runs. The experimental results are similar to those reported in the paper. Therefore, SMV performs very well compared to other state-of-the-art techniques under the same conditions.

In all 6 collections, WT10G, GOV and GOV2 are collections whose documents are crawled from the web. Unlike the three other collections, these web collections are noisy because there are many duplicates or near-duplicates, spam, documents written in foreign languages, binary data documents, etc. Some researchers (e.g., in [9]) observe that for such collections, query performance prediction is less accurate. In our experiment, we divide the 6 collections into 2 types: clean and noisy. Thus 3 web collections are classified as noisy whilst the rest are clean. For these collections, too, SMV performs better on average than each of the other methods. Table 4 gives more detailed information. We can see that on average, all performance prediction methods do better with clean collections than with noisy collections.

5. SUMMARY AND FUTURE WORK

We have presented a performance prediction method SMV by considering both score magnitude and variance at the same time. Evaluated with 6 different collections used in TREC, we find that our predictor performs better than any of the three other predictors in more data sets. Thus, we can conclude that the proposed method is very competitive.

In terms of future work, we shall focus on a few specific retrieval systems and models such as Terrier, Indri, BM25, Kullback-Leibler Divergence Language Model to further investigate the performance prediction problem. If we can treat those results from different systems/models in different ways then more accurate prediction is possible since the distribution of scores may differ from one system to another.

In a different vein, we can take more information such as certain statistics of the collection, query terms, and so on into consideration. Thus the method proposed in this paper can be used together with others for more accurate performance prediction.

6. REFERENCES

[1] J. A. Aslam and V. Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of ECIR*, pages 198–209, 2007.

[2] O. A. Shtok and D. Carmel. Predicting query performance by query-drift estimation. In *Proceedings of ICTIR*, pages 305–312, 2009.

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

[4] 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 SIGIR*, pages 512–519, 2005.

[5] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *Proceedings of SIGIR*, pages 18–24, 2004.

[6] G. Amati, C. Carpineto and G. Romano. Query difficulty ,robustness, and selective application of query expansion. In *Proceedings of ECIR*, pages 127–137, 2004.

[7] B. He and I. Ounis. Inferring query performance using preretrieval predictors. In *Proceedings of SPIRE*, pages 43–54, 2004.

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

[9] J. Pérez-Iglesias and L. Araujo. Standard deviation as a query hardness estimator. In *Proceedings of SPIRE*, pages 207–212, 2010.

[10] S. Cronen-Townsend, Y Zhou and W. Bruce Croft. Predicting query performance. In *Proceedings of SIGIR*, pages 299–306, 2002.

[11] V. Vinay, I. J. Cox, N. Millic-Frayling and K. R. Wood. On ranking the effectiveness of searches. In *Proceedings of SIGIR*, pages 398–404, 2006.

[12] Y. Zhou. Retrieval performance prediction and document quality. *PhD thesis,University of Massachusetts*, September 2007.

[13] Y. Zhou and W. Bruce Croft. Query performance prediction in web search environments. In *Proceedings of SIGIR*, pages 543–550, 2007.

[14] Y. Zhou and W. Croft. Ranking robustness: a novel framework to predict query performance. In *Proceedings of CIKM*, pages 567–574, 2006.