Predicting Query Performance in Intranet Search[*]

Craig Macdonald
University of Glasgow
Glasgow, G12 8QQ, U.K.
craigm@dcs.gla.ac.uk

Ben He
University of Glasgow
Glasgow, G12 8QQ, U.K.
ben@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, G12 8QQ, U.K.
ounis@dcs.gla.ac.uk

ABSTRACT

The issue of query performance prediction has been studied in the context of text retrieval and Web search. In this paper, we investigate this issue in an intranet environment. The collection used is a crawl of the dcs.gla.ac.uk domain, and the queries are logged from the domain search engine, which is powered by the Terrier platform. We propose an automatic evaluation methodology generating the mean average precision of each query by cross-comparing the output of diverse search engines. We measure the correlation of two pre-retrieval predictors with mean average precision, which is obtained by our proposed evaluation methodology. Results show that the predictors are very effective for 1 and 2-term queries, which are the majority of the real queries in the intranet environment.

Keywords

WWW search engines, Evaluation, Intranet Search, Query Performance Prediction

1. INTRODUCTION

The notion of predicting query difficulty refers to techniques that infer the performance of a given query, without knowing the relevance assessment information.

There has been some previous work regarding the query performance prediction issue. In [5], Cronen-Townsend proposed inferring query performance by using a clarity score, which is the divergence of a term's query language model from its collection language model. In [2], Amati et al. proposed the query difficulty notion that measures how a query term's distribution in a pseudo relevance document set diverges from randomness. In [7, 13], a set of pre-retrieval query performance predictors were proposed and studied (these predictors are pre-retrieval in the sense that their computation does not involve the use of relevance scores, unlike query clarity and query difficulty). In [9], Kwok et al. applied the support vector regression for the performance prediction. In [11] and [15], the idea of computing metrics for prediction purposes from inverse document frequency (idf) was tested.

In this work, we focus on the query performance prediction in an intranet environment. The contribution of this paper is twofold. First, we propose an automatic methodology generating the evaluation measures, e.g. average precision, precision at 10, as well as list correlation measures. Second, we compute the correlation of these evaluation measures with two pre-retrieval predictors, including the average inverse collection term frequency (AvICTF) [13] and query scope [7]. Through extensive experiments, we investigate the effectiveness of these pre-retrieval predictors in an intranet environment with real user queries.

The remainder of this paper is organised as follows: we introduce the query predictors we use in Section 2. In Section 3 we describe our new evaluation methodology that allows evaluation without relevance assessments. The experimental setting of the intranet context in which the experiments take place is described in Section 4, before describing the results we had when comparing the query performance predictors to our proposed measures in Section 5. We conclude and give future directions in Section 6.

2. QUERY PREDICTORS

In this section, we introduce the applied query performance predictors in our paper:

- Average inverse collection term frequency (AvICTF). Proposed in [7], the AvICTF predictor is generated from Kwok's idea of the *inverse collection term frequency* (ICTF) [10]:

  $$ICTF = \frac{\log_2 \frac{token_c}{F}}{ql}$$

  where F is the number of occurrences of a query term in the whole collection and $token_c$ is the number of tokens in the whole collection. ql is the query length, which is the number of unique non-stop words in the query.

  The idea of ICTF is similar to the inverse document frequency (idf): the frequency of a query term in the collection reflects its contribution in retrieval.

  AvICTF infers query performance by the average quality of the composing query terms of a query Q:

[*] ACM SIGIR'05 Query Prediction Workshop
August 19, 2005, Salvador, Brazil.

$$AvICTF = \frac{\log_2 \prod_Q \frac{token_c}{F}}{ql} \tag{1}$$

- Query scope. Proposed in [7], the query scope infers query performance by measuring the specificity of a query. It is inspired by the work of Plachouras et al. on web search [14].

  The query scope is computed as follows:

  $$-\log(N_Q/N) \tag{2}$$

  where N is the number of documents in the whole collection, and N_Q is the number of documents containing at least one of the query terms.

  According to the study in [7], query scope is effective in inferring query performance for short queries in ad-hoc text retrieval. In particular, for the single term queries, query scope can be seen as the idf and is effective in performance prediction. However, for the longer queries, as query terms co-occur in many documents, N_Q tends to be stable and becomes less effective in differentiating queries.

3. EVALUATION METHODOLOGY

In this section, we propose an automatic evaluation methodology, based on the cross-comparison of the results of several search engines, which generates several performance measures, including a mean average precision (MAP).

Traditionally, comparative evaluation of information retrieval (IR) systems has been based on three pre-requisites:

- Collection of documents
- Set of queries
- Relevance assessments for the queries

For the assessment of an IR system in an intranet setting, the collection is given, and queries can be logged from real users once the system is in place. However, relevance assessments are costly to obtain, requiring human intervention to assess each document returned from each query as relevant or non-relevant.

However, if other diverse IR systems also search the same intranet, then it is possible to compare the results from those systems, and generate an optimal ranking of documents.

3.1 Reference Ranking as an optimal ranking

We define the notion of the "Reference Ranking" (RR) which is a combination of all the IR systems that search the same collection - we assume that this combination is the "optimal ranking". We can use various techniques from the metasearch field to merge the results of all the search engines - for this paper we have used the Condorcet-fuse technique developed by Montague and Aslam in [12]. For each query, we can then compare the ranking of results produced by each search engine to the Reference Ranking.

In this paper, we selected 4 external search engines which also searched the same intranet domain - these were Google, Yahoo, MSN and Teoma. We examined but discarded the following search engines as they showed high correlations to other search engines: AlltheWeb, AltaVista (both correlated to Yahoo) and A9 (correlated to Google). These patterns closely match what is known of the commercial search engine market: Yahoo now owns both AlltheWeb and AltaVista; and A9 uses Google results as the basis for its own search results. We excluded these engines as including them would have biased the results of our evaluation.

For each query, we create the Reference Ranking from the top 100 results of each engine, (including the engine being compared to), similar to the pool of results that are assessed in TREC evaluations. The intuition behind this is that at least some documents in this pool are relevant. We then crop the Reference Ranking to the top 100 results, giving a final ranking of the most popular 100 documents for that query.

3.2 List Correlations: Spearman's Rho and Kendall's Tau

We desire a measure that allows us to directly compare the ranking produced by each engine to that of the Reference Ranking for one (identical) query. This would be straightforward if each engine produced different rankings of the same results. For example, we could use the Spearman's Rho or Kendall's Tau.

However, because each IR system ranks results differently or has indexed different documents, some results may appear in one ranking that may not appear in the other ranking. Hence, we need techniques that measure correlation even with incomplete lists. Fagin et al. devised measures that take the missing items into account [6]. Bar-Ilan suggested that these measures are only useful when the overlap between the two compared lists is large, which is not always the case when comparing the search results of different search engines [4]. She then proposed an alternative technique, where the correlation measure is computed by considering only documents that appear in both lists (i.e. the intersection lists).

In this paper, for each query we measure the correlation between the intersection lists of each IR system and the Reference Ranking using Spearman's Rho (Sp-I) and Kendall's Tau (KT-I). These measure how close the ranking of documents produced by each search engine is to the Reference Ranking. Both are defined on a scale of -1 to 1, with 1 defined as complete correlation, -1 as an inverse correlation and 0 as no correlation at all.

We then rank all queries by the mean correlation of all engines for that query.

3.3 Non-binary relevance and Precision

If we extend the notion of relevance of a document to a query to be non-binary, i.e. multi-valued, in the range [0, 1], then we can consider the Reference Ranking to define the ranking of documents that are relevant to the query. Because we assume that the Reference Ranking is optimal, then the first document in the Reference Ranking is the most relevant document to the query (intuitively because it has been placed earliest by the highest number of search engines). We define $m(d)$ to be the relevance of each document d to the query, where

$$m(d) = \begin{cases} \frac{1}{RR(d)} & \text{if Reference Ranking contains } d \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

(Note that RR(d) is the rank of document d in the Reference Ranking.) This is intuitive if we assume that the more search engines that have placed a document d higher in their rankings, then it is more likely that the document d is more relevant to the query than another document that has on average been placed lower (or not ranked at all) by the engines.

Using non-binary relevance, it is possible to calculate the counterparts of most of the normal evaluation measures [3]. Precision P at position j in a search engine ranking SE can be calculated as :

$$\begin{gathered} P_j = \tfrac{k}{j} \\ \text{with the greatest } k \text{ such that} \\ \textstyle\sum_{a=1}^{k} m(RR(a)) <= \sum_{b=1}^{j} m(SE(b)) \end{gathered} \tag{4}$$

where $SE(i)$ is the document at rank i in search engine ranking SE, and similarly $RR(i)$ is the document at rank i in the Reference Ranking. Note that $k \leq j$ since the Reference Ranking is the optimal ordering - i.e. $m(d)$ is strictly descending on RR.

The definition of mean average precision (MAP) easily follows:

$$MAP = \frac{1}{min\{|SE|, |RR|\}} \sum_i P_i \tag{5}$$

In this paper, we use Precision at 10 and MAP to measure performance of the search engines.

4. EXPERIMENTAL SETTING

In this section, we describe the existing search engine on the domain, with how and when data was gathered for the experiments. The aim of our experiments is to measure the correlation between our proposed measures and our existing query performance prediction measures over different subsets of queries.

4.1 Existing intranet search engine

For the last year, we have been providing a search engine on the www.dcs.gla.ac.uk website. The dcs.gla.ac.uk domain consists of several websites of the Department of Computing Science at the University of Glasgow. The contained web pages fall into several categories, including personal home pages, research group related material, teaching, student recruitment and administrative pages. The search engine is used by users who have an information need while using these websites.

The intranet search engine is based on Terrier. Terrier is a platform for the rapid development of large-scale IR applications[1]. Terrier has various weighting models and retrieval approaches, including the Divergence from Randomness (DFR) models [1]. The intranet search engine uses the DFR weighting model PL2, given by:

$$\begin{aligned} score(d, Q) \quad = \quad & \sum_{t \in Q} \frac{qtf}{tfn + 1} \big(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \\ & \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn) \big) \end{aligned} \tag{6}$$

where $score(d, Q)$ is the relevance score of a document d for a query Q. t is a query term in Q. λ is the mean and variance of a Poisson distribution. qtf is the query term frequency. The normalised term frequency tfn is given by the *normalisation 2*:

$$tfn = tf \cdot \log_2 (1 + c \cdot \frac{avg\_l}{l}), (c > 0) \tag{7}$$

where l is the document length and avg_l is the average document length in the whole collection. tf is the original term frequency. c is the free parameter of the normalisation method. For this paper, we used a value of 1.18 determined automatically using the technique described by He & Ounis in [8].

We have extended the pure content search used by Terrier with tested techniques suitable for use in a Web IR setting [13]:

- Firstly, we extend documents by adding the anchor text of their incoming hyperlinks to the body of the document.
- If a term occurs in the TITLE or H1 tags of a document, or if it occurs in the anchor text of the document's incoming hyperlinks, we boost the score of the term in that document by 10%.
- Finally, we use a technique called URL scoring to rerank the top 50 documents retrieved using content and anchor text retrieval techniques, using the score:

  $$score_i = s_i \times \frac{1}{\log_2 (urlpath\_len_i + 1)} \tag{8}$$

  where s_i is the score assigned to document d_i, and $urlpath\_len_i$ is the length in characters of the URL path of d_i.

Notice that the external search engines we are using for comparison all appear to be using a boolean filtering of results, i.e. they return documents that only contain all of the query terms. To ensure a suitable setting for comparison in our experiments, Terrier has been likewise configured to only return documents that contain all of the query terms.

4.2 Experimental Setup

For these experiments we will be using the collection of documents obtained by crawling our intranet. We obtained crawls of the dcs.gla.ac.uk domain of the world wide web on the 18th April 2005 using our own web crawler called Labrador[2]. Terrier indexes this collection by removing English stop-words then applying the first two steps of Porter's Stemmer. This provides an index of 49,354 documents with 339,401 unique terms, including anchor texts.

By logging the queries submitted to the intranet search engine, we can gather queries that are representative of real users information need on our intranet. We logged the queries submitted to our search engine over a 32-week period, from 19th December 2004 to 18th April 2005. We aggregated and then manually removed any queries that any of the search engines would not be able to handle. This gave us 1702 unique queries with which to compare the performance of Terrier to the external search engines. As shown in Table 1, the average query length is very low, typical with a Web IR setting.

We built several subsets of queries: 1, 2, 3 and 4-term queries; phrasal queries; and queries which were about people. This allows us to assess the correlation of the query performance predictors over the different query subsets.

[1] See http://ir.dcs.gla.ac.uk/terrier/

[2] See http://www.dcs.gla.ac.uk/∼craigm/labrador/

Number of queries	1702
Average Query Length	1.9
1-term queries	727
2-term queries	583
3-term queries	234
4-term queries	82
people queries	134
phrasal queries	35

Table 1: Query Statistics

Engine	Mean Results Size
Teoma	43.4974
MSN	40.5144
Yahoo	45.1311
Google	52.3492
Terrier	43.5867

Table 2: Average number of results over all queries

Using our new methodology, we can determine the performance of the search engines for the set of queries, without any need for relevance assessments, by combining the search results of many IR systems that search the same collection. We submitted all queries to the search engines in the period 18th to 22nd April 2005, to ensure that the results obtained were as close to the timescale of our own crawl, minimising any error in our experiments caused by the indices of the external search engines changing. For each external engine, we appended `site:dcs.gla.ac.uk` to the query, so that only results from the same domain as our own search engine were returned. As can be seen in Table 2, the average number of results returned by each search engine is very similar, so it can be assumed that the search engines each have good coverage of the domain. However, we note that Google returns more documents than all the other search engines. This is maybe due to its comparatively high coverage of the web, allowing it to use the additional anchor text information to match more intranet documents for each query.

For different subsets of the queries, we produce rankings of queries based on the evaluation measures, and then compare these to the rankings of queries based on the predictors. We compute the Spearman's Rho of the predictors with the evaluation measures, including mean average precision (MAP), Kendall's tau of intersection (KT-I) and Spearman's Rho of intersection (Sp-I).

5. DISCUSSION OF RESULTS

In this section, we discuss the results. Table 3 contains the obtained mean average precision (MAP), Precision at 10 (P@10), Kendall's Tau of intersection (KT-I) and Spearman's Rho of intersection (Sp-I) correlations for each search engine. As shown in the table, the two measures of intersection are not necessarily correlated with MAP.

However, there appears to be a correlation between the average number of returned results and the MAP achieved by a search engine. This can be explained by the fact that the proposed MAP measure has a recall component - the more results a search engine returns, the more probable that the achieved MAP is higher. Returning a different number of results could be linked to domain coverage, or to external search engines using anchor text from elsewhere on the web to match additional documents in the intranet.

Search Engine	MAP	P@10	KT-I	Sp-I
Teoma	0.4092	0.3583	0.3727	0.4656
MSN	0.3482	0.3497	0.3922	0.4918
Yahoo	0.4018	0.4380	0.3966	0.4959
Google	0.5121	0.5500	0.4722	0.5831
Terrier	0.3841	0.4378	0.4245	0.5229

Table 3: Search engine performance achieved over all queries.

On the other hand, there appears to be no correlation between the KT-I and Sp-I measures and the number of results returned by each search engine, therefore these measures may be more reliable for measuring the retrieval performance of search engines.

Further investigation is required to check whether the same findings would be obtained on other intranet domains or collections. In particular, whether the correlation measures Sp-I and KT-I would agree more with MAP when the number of results returned by each search engine was fixed. This would limit the influence of the recall component in the MAP definition.

Table 4 presents the obtained Spearman's correlation of the predictors with the evaluation measures for 1, 2, 3, 4-term queries and all the queries, respectively. According to the results:

- AvICTF is shown to be effective in intranet search. It has a high correlation with MAP, particularly for single-term queries. It is also noticed that the correlation decreases when query length increases.

- We have a similar observation for the query scope. Its correlation with MAP decreases when query length increases. However, the correlation with MAP seems to be very sensitive to the query length. In particular, for the 4-term queries, the correlation seems to be random, i.e. 0.0208. This confirms our hypothesis in Section 2 that when the query becomes longer, query scope tends to be stable, while its differentiating power of query performance decreases.

- The predictors are not highly correlated with the intersection of the search engines with the Reference Ranking (the KT-I and Sp-I measures). Our explanation is that the predictors are based on the frequency of the queries terms in the collection. The underlying hypothesis is that the larger the frequency is, the poorer the query performs. On the other hand, when the frequency is small, the search engines agree on a small set of documents that are highly relevant, while they have less agreements on other documents. Therefore, the correlation of the predictors with KT-I and Sp-I is relatively low. This explanation is supported by Table 3. The KT-I and Sp-I are not necessarily correlated with MAP.

We also computed the correlation of the predictors with MAP for subsets of the queries, including 134 queries searching for persons (see Table 5) and 35 phrasal queries (see

	MAP	KT-I	Sp-I
1-term queries			
AvICTF	0.7109	0.1555	0.1136
Query Scope	0.7329	0.1772	0.1320
2-term queries			
AvICTF	0.4557	0.3006	0.2825
Query Scope	0.2808	0.2590	0.2558
3-term queries			
AvICTF	0.4117	0.1786	0.1739
Query Scope	0.1702	0.0353	0.0456
4-term queries			
AvICTF	0.3288	0.5071	0.4782
Query Scope	0.0208	0.3082	0.2970
All queries			
AvICTF	0.4389	0.1547	0.1367
Query Scope	0.2991	0.0882	0.0795

Table 4: Spearman's Rho of the predictors with the evaluation measures for 1, 2, 3, 4-term queries and all the queries, respectively.

	MAP	KT-I	Sp-I
AvICTF	0.1832	-0.0869	-0.1210
Query Scope	-0.0869	-0.1947	-0.2356

Table 5: Spearman's Rho of the predictors with the evaluation measures for queries searching for persons.

Table 6). According to the results, apart from AvICTF which has a decent correlation with MAP for the persons queries, the predictors have relatively weak correlation with the evaluation measures. As the applied predictors do not account for the position of query terms in documents, it is expectable that they do not perform well for the phrasal queries. For the queries searching for persons, our explanation is that these queries search for the people's home pages. The number of relevant documents are very few and stable. Therefore, the query performance is relatively stable, which leads to weak correlation of the predictors with query performance. We have computed the variance of MAP for all the queries and for the queries searching for persons, respectively. The obtained variance values are 0.0614 for the former and 0.0292 for the latter, which support our explanation that queries searching for persons tend to have stable performance.

Overall, AvICTF and query scope seem to be effective for 1 and 2-term queries. Since this is an intranet environment, where most real queries consist of only 1 or 2 terms (see Table 1), we can conclude that AvICTF and query scope can be applied in intranet search as accurate query performance predictors.

	MAP	KT-I	Sp-I
AvICTF	0.0652	0.0071	0.0553
Query Scope	-0.0762	-0.0230	0.0345

Table 6: Spearman's Rho of the predictors with the evaluation measures for phrasal queries.

6. CONCLUSIONS AND FUTURE WORK

The merits of this paper are twofold. We have proposed an evaluation methodology based on the cross-comparison of the output of diverse search engines. We also studied the query performance prediction in intranet search. Results show that the applied predictors, including the average inverse collection term frequency (AvICTF) and the query scope, are overall effective in intranet search. Their effectiveness is particularly high for 1 and 2-term queries, but decreases as query length increases. Moreover, query scope seems to be extremely sensitive to the query length. Finally, we find relatively weak correlation of the predictors with the intersection measures, i.e. KT-I and Sp-I.

In the future, we intend to assess the effectiveness of our evaluation methodology by comparison in settings where relevance assessments are available, for example using TREC submissions and relevance assessments. Furthermore, we would like to study applications of query performance predictors in intranet search setting, with a view to developing techniques to improve retrieval performance. For example, applying appropriate retrieval approaches when a query is predicted to perform poorly. In addition, we need to develop predictors that are effective in predicting performance for home page and known item retrieval tasks, as well as for phrasal queries.

7. REFERENCES

[1] G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness.* PhD Thesis, Department of Computing Science, University of Glasgow, 2003.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Advances in Information Retrieval, Proceedings of the 26th European Conference on IR Research, ECIR 2004*, pages 127 – 137, Sunderland UK, April 2004.

[3] G. Amati and F. Crestani. Probabilistic learning by uncertainty sampling with non-binary relevance. In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval: techniques and applications*, pages 299–313. Physica Verlag, Heidelberg, Germany, 2000.

[4] J. Bar-Ilan. Comparing rankings of search results on the web. *Information Processing and Management.* In press. doi:10.1016/j.physletb.2003.10.071

[5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299 – 306, Tampere, Finland, 2002.

[6] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 2003*, pages 28 – 36, Baltimore, MD, 2003.

[7] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proceedings of the SPIRE 2004*, pages 43 – 54, Padova, Italy, October 2004.

[8] B. He and I. Ounis. Tuning Term Frequency Normalisation for BM25 and DFR Models. In *Proceedings of the 27th European Conference on Information Retrieval (ECIR'05)"*, pages 200 – 214, Santiago de Compostela, Spain, March, 2005.

[9] K. Kwok, L. Grunfeld, H. Sun, P. Deng, and N. Dinstl. Trec 2004 robust track experiments using pircs. In *Proceedings of The Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD, 2004.

[10] K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187 – 195, Zurich, Switzerland, 1996.

[11] L. Si and J. Callan. Using sampled data and regression to merge search engine results. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–26, Tampere, Finland, 2002.

[12] M. Montague and J. A. Aslam. Condorcet fusion for improved retrieval. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, New York, NY, USA, 2002. ACM Press.

[13] V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with terrier. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD, 2004.

[14] V. Plachouras, I. Ounis, G. Amati, and C. J. V. Rijsbergen. University of Glasgow at the Web Track: Dynamic application of hyperlink analysis using the query scope. In *Proceedings of the Twelth Text REtrieval Conference (TREC 2003)*, pages 248 – 254, Gaithersburg, MD, 2003.

[15] F. Scholer, H. Williams, and A. Turpin. Query association surrogates for web search. In *Journal of the American Society for Information Science and Technology*, number 55(7), pages 637–650, 2004.