Predicting Query Performance Using Query, Result, and User Interaction Features

Qi Guo[1], Ryen W. White[2], Susan T. Dumais[2], Jue Wang[2], and Blake Anderson[2]
[1] Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA
[2] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA
qguo3@emory.edu, {ryenw,sdumais,juewan,blakean}@microsoft.com

ABSTRACT
The high cost of search engine evaluation makes techniques for accurately predicting engine effectiveness valuable. In this paper we present a study in which we use features of the query, search results, and user interaction with the search results to predict query performance. We establish which features are most useful, study the effect of different classes of features, and examine the effect of query frequency on our predictions. Our findings show that performance predictions using result and interaction features are substantially better than those obtained using only query features. Such results can support automated search engine evaluation methods and new query processing capabilities.

Keywords
Query performance prediction

1. INTRODUCTION
The ability to accurately evaluate the quality of search engines is important in improving retrieval algorithms and in identifying queries or query classes that are particularly challenging. The query evaluation process usually involves time-consuming and expensive human judgments or user studies that only cover a small fraction of queries that search engines receive. Developing an accurate automated method to estimate search engine performance would reduce these costs and facilitate more efficient evaluation of new retrieval models, ranking algorithms, or presentation techniques. Prior work by Cronen-Townsend and Croft [3], He and Ounis [9], and Hauff et al. [8] has developed automatic methods for predicting query difficulty (i.e., how good the search results are for a query). Measures such as *clarity* [3] that compare the distribution of terms in search results to their distribution in the entire collection have shown some promise. Also, some prior work has modeled user interaction to predict search result preferences [1] or search goal success [7]. However, little work has been done in predicting query performance using searcher interactions. Carterette and Jones [2] used click-through behavior to evaluate the quality of search advertising results, but they did not study other user interaction features, and their focus was on search advertising not general Web search.

Our work differs from the previous research in that we investigate a richer set of features derived from user queries, search results, and user interactions with search results. Also, our experiments are conducted on datasets derived from the logs of a large-scale commercial Web search engine, rather than standard information retrieval test collections, which are small-scale and less representative of the diversity of Web search scenarios. The primary contributions of this paper are: (i) investigate a novel and richer set of interaction features in predicting query performance, (ii) determine which features and combinations of features are most important in the prediction task, and (iii) understand how prediction performance varies with query popularity. Accurately predicting query performance allows search engine designers to tailor the search experience for particular queries or query classes, e.g., by applying different query processing or presentation techniques for queries that are predicted to be of poor quality.

In Section 2 we describe our prediction technique and how we evaluate its performance. Section 3 presents the experimental findings. We conclude in Section 4 with some study implications.

RIAO'10, 2010, Paris, France.

2. PREDICTING QUERY PERFORMANCE
To predict query performance we must first define how we measure query performance in our study. We use discounted cumulative gain (DCG) [6], an established measure of retrieval effectiveness that uses graded relevance assessment and measures the usefulness, or *gain*, of a document based on its position in the result list. In this paper we predict DCG at rank position three ($DCG@3$), so as to capture the relevance of the top three search results. In our computation of DCG, we use a logarithmic discount for position and an exponential gain for five levels of relevance: *bad*, *fair*, *good*, *excellent*, and *perfect*. Specifically, for a query:

$$DCG@3 = \sum_{i=1}^{3}(2^{rel_i} - 1)/log_2(1+i),$$
$$\text{where } rel_i \in \{0,2,3,4,5\}.$$

This results in a minimum $DCG@3$ of 0 and a maximum $DCG@3$ of 66.1. We aim to develop a model which lets us automatically predict the DCG score for each query based on feature weights learned during training.

We now describe the features used to make query performance predictions, the data sources, the model, and the metrics that determine predictor performance.

2.1 Features
The features were divided into three classes: *Query*, from logs containing the stream of incoming queries received by a search engine; *Results*, from parsing the search engine result page (SERP) for those queries, and; *Interaction*, from logs that include those queries' SERP interactions (e.g., clicks on SERP links) and post-SERP behavior (e.g., search engine switching, where users voluntarily transition between engines). All logs were collected during a one-week period in July 2009 from the Bing search engine. Queries and SERP clicks were captured using server-side

logging. Post-SERP interactions were captured using a widely-distributed browser toolbar. Earlier work has shown that switching behavior can be indicative of dissatisfaction with search results [11], therefore, we include switching behavior in our interaction feature set. Table 1 summarizes the features for each query.

Table 1. Features used in study.

Feature	Feature description
Query class	
QueryLength	Number of characters in the query
QueryWordLength	Number of words in the query
HasURLFragment	True if the query contains a URL fragment (e.g., "http://", ".com")
HasSpellCorrection	True if search engine spelling correction is offered for query
HasAlteration	True if query is automatically modified by engine (e.g., stemming)
Results class	
AvgNumAds	Average number of advertisements shown on the query's SERP
AvgNumQuerySuggestions	Average number of query suggestions shown on the query's SERP
AvgNumResults	Average number of total search results for the query
HasDefinitive	True if a single best result for the query is included in the result set (usually for navigational queries)
MaxBM25F	Maximum BM25F retrieval score for query across all search results
MaxBM25FNorm	MaxBM25F for query, normalized based on top MaxBM25F score
Interaction class	
QueryCount	Number of query occurrences
ClickCount	Number of SERP clicks for query
SATCount	Number of SERP clicks for query followed by a dwell time exceeding 30 secs. or session termination
AvgSerpDwell	Avg. SERP dwell time for query
AvgClickDwell	Avg. dwell time after SERP click
CTRate	*ClickCount*/*QueryCount*
SATRate	*SATCount*/*ClickCount*
AvgClickPos	Avg. SERP click position for query
AvgNumClicks	Avg. num. SERP clicks for query
AbandonmentRate	Fraction of times query issued and has no SERP click
PaginationRate	Fraction of times query issued and next page of results requested
SwitchCount	Number of query occurrences followed by an engine switch, with the same query issued on both pre- and post-switch engines
SwitchRate	*SwitchCount*/*QueryCount*

In our feature computations, we define a *search session* as a group of consecutive search events (queries or search result clicks) terminated by a 30-minute period of inactivity. Similar thresholds have been used to demarcate log sessions [10]. A *switching query pair* is two consecutive (and identical) queries issued on two different search engines within the same search session. Dwell time on the SERP is computed based on the time difference between the result page loading and a user action such as a click or re-query. Dwell time following a SERP click is defined as the amount of time spent away from the SERP until the next click. Previous work has defined a *satisfied* SERP click as one where a click on a search result is followed by a non-SERP dwell of 30 seconds or more [4]. We use that definition to compute *SATCount*.

2.2 Data

We used a set of 2,834 queries obtained by randomly sampling the query logs of the major US commercial search engine used in this study. The set comprised a mixture of common and rare queries for which we had human relevance judgments and could generate all the features listed in Table 1. The top ten results for each query were downloaded and parsed to generate the *Results* features. In addition, those results were judged for relevance by human assessors using the five-point scale described earlier. These explicit judgments are used as the ground truth to generate the DCG values that we aim to predict. The *Query* and *Interaction* features were obtained from logs from the same search engine that captured both interactions with the engine and browser navigation patterns (so that we could detect search engine switching events). Features and judgments were all obtained during the same time period in July 2009. We used 60% of the 2,834 queries for training, 20% for validation and 20% for testing, and performed five-fold cross validation to improve result reliability.

2.3 Model

We used Multiple Additive Regression Trees (MART) [5] to train a regression model to predict $DCG@3$. MART uses gradient tree boosting methods for regression and classification. Advantages of MART include model interpretability (e.g., a ranked list of important features is generated), facility for rapid training and testing, and robustness against noisy labels and missing values.

2.4 Metrics

To evaluate model predictive performance we used Pearson's correlation (R) and mean absolute error (MAE). Correlation measures the strength of association between predicted DCG and actual DCG on a scale from −1 to 1, with one indicating a perfect correlation and zero indicating no correlation. MAE is the mean of the absolute deviations between the actual DCG and predicted DCG. The ideal value is zero, with larger values showing more errors. Although MAE could range from 0 to 66.1, we normalized it to range from 0 to 1 to assist with interpreting the findings.

3. EXPERIMENTAL RESULTS

3.1 Overall

Using all available features, we can effectively predict DCG using the learned model – R=0.699, MAE=0.160. The model trained and tested using *Query*, *Results*, and *Interaction* features is referred to as the *full* model in the remainder of the paper. Figure 1 shows predicted DCG versus actual DCG for all 2,834 queries in our set. The predicted and actual DCG values are normalized to range from 0 to 1. Since our experimental methodology used five-fold cross validation, each query was a test query in exactly one fold.

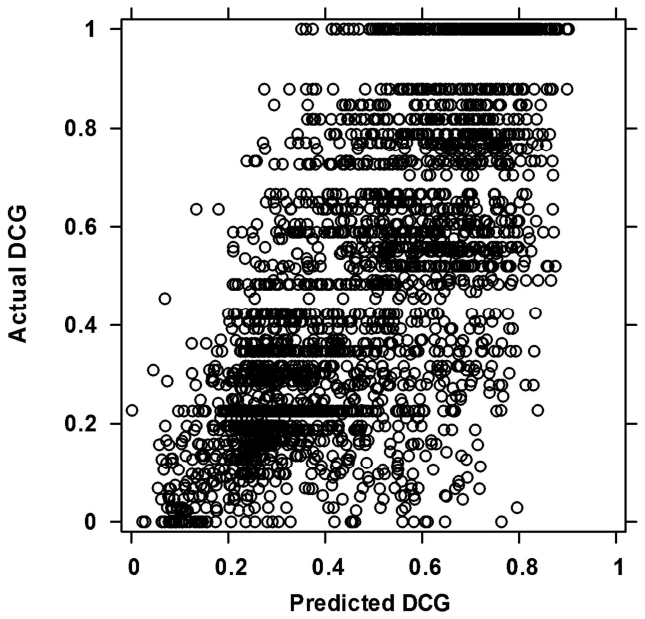

Figure 1. Predicted DCG versus actual DCG (full model).

The distribution of the predicted DCG scores demonstrates that the correlation appears sensible throughout the range.

Table 2 shows a list of the most important features in one of the five experimental runs ranked by importance to the full model relative to the most important feature, *AvgClickPos*. The feature ordering shown in the table is typical of that found in other runs. The table shows that many *Interaction* features are important, including average click position, average number of clicks, abandonment rate, CTRate, pagination rate, and search engine switch rate. Some *Results* features are also important, including average number of shown query suggestions, BM25F score, and the average number of search result advertisements shown. *Query* features, including average query word length also contribute, although to a lesser extent.

Table 2. Feature contributions.

Feature Name	Feature Type	Importance
AvgClickPos	Interaction	1
AvgNumClicks	Interaction	0.542
AvgNumQuerySuggestions	Results	0.294
AbandonmentRate	Interaction	0.270
BM25F	Results	0.268
ClickCount	Interaction	0.253
BM25FNorm	Results	0.248
CTRate	Interaction	0.245
PaginationRate	Interaction	0.236
AvgNumAds	Results	0.193
QueryWordLength	Query	0.192
AvgClickDwell	Interaction	0.179
SwitchRate	Interaction	0.178

We also examined cases where the model predictions disagree with the human judgments. Several instances involved cases where the top results were presented in novel ways (e.g., richer search result captions). Thus novel presentation techniques may influence of some of the interaction features. Space limitations preclude the presentation of more detailed failure analyses.

3.2 Feature Combinations

To further understand the contribution of different feature groups, we trained six regression models, using all combinations of the three feature classes. The results are shown in Table 3. All significant differences with paired t-tests between the sub-models and the full model (in terms of predicted DCG and MAE) are highlighted in the table (* = p < .05, ** = p < .01).

Table 3. Feature sets used in feature ablation and their associated performance scores.

Feature set	R	MAE
Query + Results + Interaction (full model)	0.699	0.154
Results + Interaction	0.698	0.160
Query + Interaction	0.678	0.164
Interaction only	0.667 *	0.166 *
Query + Results	0.556 **	0.193 **
Result only	0.522 **	0.200 **
Query only	0.323 **	0.228 **

The analysis shows that the query only features perform poorly, and the addition of the query features does not add much to the performance of the *Results* or *Interaction* features. *Results* features perform reasonably well. This condition represents a text-matching baseline, which includes BM25 but is somewhat better. *Interaction* features alone produce a performance that is close to that of all features combined. It seems that there is strong predictive signal in interaction behavior.

3.3 Query Breakdown

Our analysis so far has considered only the performance across all queries. Given the importance of user interaction features we also considered the effect of query frequency (and thus the amount of interaction data) on performance. We ranked queries in descending order by *QueryCount*. We then started from the top of the list and divided the queries into equally-sized bins of 50 queries each with no query overlap. This procedure generated 57 bins, of varying frequency, across the 2,834 queries. Within each bin, we computed the correlation (R) between the predicted DCG score from our full model and the actual DCG scores obtained from human judgments. A linear regression across all R values revealed that, for the set of queries that we examined, there was no apparent relationship between the query frequency (and thus the amount of behavioral data) and the accuracy of our predictions (R^2=.008).

4. CONCLUSIONS AND FUTURE WORK

In this paper we study automatically predicting the quality of Web search engines by evaluating different sources of evidence derived from the queries, results, and user interaction logs. We showed a strong correlation (~R=0.7) between predicted DCG and human relevance judgments when all features are used. The findings also show that how users interact with search results provides a strong

signal of the quality of search results for queries, adding substantially to results obtained using *Query* and *Results* features.

We believe that the ability to accurately predict query performance can be used to support a variety of search-related enhancements. For example, it enables search engine designers to apply different query processing, ranking, or presentation methods for queries of different quality, to sample queries of different quality, and to identify poor performing queries. Further research is required to understand the role of other features, effects related to the nature of the document collection used, and the impact of search engine settings on prediction effectiveness.

5. REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. Learning user interaction models for predicting web search result preferences. In Proc. SIGIR, 3-10 (2006).

[2] Carterette, B. and Jones, R. Evaluating search engines by modeling the relationship between relevance and clicks. In Proc. NIPS, 217-224 (2007).

[3] Cronen-Townsend, S., Zhou, Y., and Croft, W.B. Predicting query performance. In Proc. SIGIR, 299-306 (2002).

[4] Fox, S., Karnawat, K., Mydland, M., Dumais, S.T., and White, T. Evaluating implicit measures to improve the search experience. ACM TOIS, 23(2): 147-168 (2005).

[5] Friedman, J.H., Hastie, T., and Tibshirani, R. Additive logistic regression: A statistical view of boosting. Technical Report, Department of Statistics, Stanford University (1998).

[6] Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. ACM TOIS, 20(4): 422-446 (2002).

[7] Hassan, A., Jones, R., and Klinkner, K. Beyond DCG: User behavior as a predictor of a successful search. In Proc. WSDM, 221-230 (2010).

[8] Hauff, C., Murdock, V., and Baeza-Yates, R. Improved query difficulty prediction for the web. In Proc. CIKM, 439-448 (2008).

[9] He, B. and Ounis, I. Inferring query performance using pre-retrieval predictors. In Proc. SPIRE, 43-54 (2004).

[10] White, R.W. and Drucker, S.M. Investigating behavioral variability in Web search. In Proc. WWW, 21-30 (2007).

[11] White, R.W. and Dumais, S. Characterizing and predicting search engine switching behavior. In Proc. CIKM, 87-96 (2009).