Predicting Query Difficulty on the Web by Learning Visual Clues

Eric C. Jensen, Steven M. Beitzel,
David Grossman, Ophir Frieder
Illinois Institute of Technology Information Retrieval Laboratory
Chicago, IL 60616
{ej,steve,grossman,frieder}@ir.iit.edu

Abdur Chowdhury
Search & Navigation Group
America Online, Inc.
Dulles, VA 20166
cabdur@aol.com

ABSTRACT

We describe a method for predicting query difficulty in a precision-oriented web search task. Our approach uses visual features from retrieved surrogate document representations (titles, snippets, etc.) to predict retrieval effectiveness for a query. By training a supervised machine learning algorithm with manually evaluated queries, visual clues indicative of relevance are discovered. We show that this approach has a moderate correlation of 0.57 with precision at 10 scores from manual relevance judgments of the top ten documents retrieved by ten web search engines over 896 queries. Our findings indicate that difficulty predictors which have been successful in recall-oriented ad-hoc search, such as clarity metrics, are not nearly as correlated with engine performance in precision-oriented tasks such as this, yielding a maximum correlation of 0.3. Additionally, relying only on visual clues avoids the need for collection statistics that are required by these prior approaches. This enables our approach to be employed in environments where these statistics are unavailable or costly to retrieve, such as metasearch.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Query Difficulty, Web Search

1. INTRODUCTION

Web search presents a new dimension to the problem of predicting query difficulty, in that difficulty is closely tied to users' perceptions of what is relevant and what is not. When examining web search results, users are not likely to read every document returned to them. More likely, they will make an initial determination of relevance based on the information contained in the brief document representations displayed by the search engine, called surrogates. Surrogates are designed to contain enough information for a human searcher to determine whether or not the represented document is relevant to the search in question. The surrogates for most web search engines include a title, a URL, a short snippet (keyword in context), and the rank. These surrogates allow the user to make a preliminary determination of what is likely to be most relevant, and to click on and examine only those documents in greater detail.

Given that users' perceptions of relevance are based on a visual examination of the provided surrogates, it stands to reason that visual features extracted from these surrogates might be good predictors of a query's difficulty. Other domains in information science, such as the Document Understanding Conference (DUC) have used visual clues for pseudo-evaluation. The goal of this study is to use the visual clues present in surrogates as a source of information for predicting the overall difficulty of a query.

2. PRIOR WORK

Seminal work in predicting query difficulty was done by Cronen-Townsend & others in the development of the "clarity" score [2]. Clarity scores predict a query's difficulty by measuring its relative ambiguity. In support of this, the original study found moderately strong correlations between clarity scores and mean average precision scores for ad-hoc queries from several TREC years. This approach has only been applied to TREC datasets that use a recall-oriented evaluation via deep-pooling. As such, it is unclear how well metrics like the clarity score will perform when applied to a high-precision task such as web search. He and Ounis examined techniques for predicting query performance prior to retrieval, in the hope that such a method would be more applicable to practical situations where query processing time is critical [3]. Their results showed that some pre-retrieval features obtainable from the query (such as a simplified version of the clarity score) have a significant correlation with mean average precision on TREC tasks. To our knowledge, no study exists that tries to apply these techniques in a precision-oriented environment.

3. METHODOLOGY

We propose a method for predicting the difficulty of a query using supervised machine learning on a combination of visual features from document surrogates. First, a set of queries are evaluated manually by assessors to determine each engines' score for every query. Each query is submitted to the engines of interest and ranked lists (top 10, 20, etc.) of document surrogates are retrieved for each engine. For each surrogate, as many visual features as possible are collected, i.e. percentage of character n-grams from the query appearing in the title, snippet and URL, percentage of query terms and phrases appearing, average distance between those terms, etc. These features are then aggregated across all surrogates retrieved from every engine to find per-query features such as the average percentage of character n-grams from the query appearing in titles, etc. A supervised regression algorithm is then trained to predict the average score for any given query using the manually evaluated queries and their corresponding values for these features as a training set.

4. EXPERIMENTATION & RESULTS

To evaluate our prediction methodology, we use an available precision-oriented test collection [1]. This collection consists of the top ten search results from ten web search services, manually evaluated for binary relevance over a sample of 896 queries randomly sampled form an AOL™ Search query log. The ten web search engines included are Google™, Yahoo™, Wisenut™, Teoma™, Altavista™, AllTheWeb™, Lycos™, Gigablast™, MSN™, and the MSN™ TechPreview (hereafter anonymized and referred to in no particular order as E1-10). In this collection, the

SIGIR 2005, August 15--19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

average number of unique results for a given query across all engines is 43 and the mean percentage of results in common between any two engines is 37%. We designated 2/3 of the 896 queries as the training set and 1/3 as the testing set and held these sets constant across experiments.

To examine the utility of query-based predictors from prior studies, we used collection statistics from the TREC WT10g web collection to estimate the standard deviation of inverse document frequencies of each query term, ratio of maximum query term IDF to minimum, and simplified clarity score [3]. Although they rely on collection statistics, each of these are pre-retrieval predictors that could be applied in a practical web search environment. The correlations of each of these predictors with mean P@10 across all 10 engines for all 896 queries are shown in Table 1. As in prior work, we use the non-parametric Spearman rank correlation coefficient throughout our experiments. To provide a baseline for our supervised methodology, we also applied the learning algorithm to these predictors to learn the optimal linear function of each of them independently and all of them combined (*qDifficulty*) on our training set (flipping the sign of the correlation due to negative learned weights). While simple clarity is not affected substantially, it is clear that learning the appropriate weight and intercept for the IDF-based predictors vastly improves performance.

To evaluate the effectiveness of our visual predictors, we extracted 31 visual features from the surrogate representation of each result document, and learned the relative weights of these features to predict average P@10 across engines using the SMO support vector machine regression implementation in WEKA (http://www.cs.waikato.ac.nz/ml/weka), terming this *vDifficulty*. As is evident from Table 1, this has a substantially stronger correlation than any of our query-based feature baselines. Somewhat surprisingly, learning on the union of all visual and query-based features does not improve performance. To verify that aggregating these surrogate features was appropriate, we also experimented with using the SMO algorithm to learn binary classifications of each surrogate as either relevant or not relevant and calculating the predicted P@10 scores from those classifications. Also surprisingly, this yielded a correlation of only 0.54, underperforming the regression over the aggregated features.

To analyze the reliability of our predictions on each engine individually, we performed the SMO regression on each engine independently, using features aggregated over only that engine's surrogates. To examine what differences in feature weights might exist between engines, we included the entire set of visual and query-based features in this learning. The correlations with each engines' P@10 and the largest magnitude feature weight from the visual features versus the query-based features are shown in Table 2. The top visual feature was always the average percentage of character n-grams in surrogate titles, while the top query-based feature was σ_{idf} for E1 and E6 and the ratio of minimum to maximum IDF for all others. Although the correlations fluctuate somewhat from engine to engine, they are relatively the same level. None of the individual engines' correlations are quite as large as that of the entire set combined, perhaps due to the reduction in the number of surrogates features are aggregated over.

Table 1: Predictors and Mean P@10 Correlations, All Engines

Predictor	Over All 896 No Learning	Over Test 300 Using SMO reg
σ_{idf}	-0.0544	0.2458
$\frac{idf_{\max}}{idf_{\min}}$	-0.1788	0.2985
Simp. Clarity Score	0.2458	0.2625
qDifficulty	NA	0.2457
vDifficulty	NA	**0.5735**
vDifficulty and qDifficulty	NA	0.5633

Table 2: Learning using vDifficulty and qDifficulty Combined Features for Each Engine

	Correlation Over Test 300 Using SMOreg	Avg. % n-grams in Title Weight	Best Query-Based Weight
E1	0.5049	0.3805	0.1690
E2	0.4928	0.5074	-0.3664
E3	0.5413	0.5234	-0.2885
E4	0.5641	0.4157	-0.4290
E5	0.5072	0.4220	-0.3886
E6	0.5599	0.4484	0.1367
E7	0.5303	0.6160	-0.3086
E8	0.5399	0.6824	-0.1212
E9	0.5722	0.5909	-0.1401
E10	0.5055	0.7707	-0.2714

5. CONCLUSION

We have developed a method for predicting query difficulty for the high-precision task of web search by using visual features from surrogate document representations in search results. We show that this approach has a moderate correlation of 0.57 with scores from manual relevance judgments of the top ten documents retrieved by ten web search engines over 896 queries, and outperforms the best query-based approach (0.29). In addition, we have found that predicting relevance for each result and calculating P@10 over these predictions slightly underperforms the regression over the means of our visual features. Also, we have found that prediction effectiveness and relative importance of query-based vs. visual features varies somewhat across search engines, but visual features are almost always much more effective as predictors of difficulty.

6. REFERENCES

[1] E. C. Jensen, et al. *A Framework for Determining Necessary Query Set Sizes to Evaluate Web Search Effectiveness.* In Proceedings of WWW'05.

[2] S. Cronen-Townsend, et al. *Predicting Query Performance*. In Proceedings of SIGIR'02.

[3] B. He and I. Ounis. *Inferring Query Performance Using Pre-Retrieval Predictors.* In Proceedings of SPIRE'04.