Estimating Query Performance using Class Predictions

Kevyn Collins-Thompson
Microsoft Research
1 Microsoft Way
Redmond, WA USA 98052
kevynct@microsoft.com

Paul N. Bennett
Microsoft Research
1 Microsoft Way
Redmond, WA USA 98052
paul.n.bennett@microsoft.com

ABSTRACT

We investigate using topic prediction data, as a summary of document content, to compute measures of search result quality. Unlike existing quality measures such as query clarity that require the entire content of the top-ranked results, class-based statistics can be computed efficiently online, because class information is compact enough to precompute and store in the index. In an empirical study we compare the performance of class-based statistics to their language-model counterparts for predicting two measures: query difficulty and expansion risk. Our findings suggest that using class predictions can offer comparable performance to full language models while reducing computation overhead.

1. INTRODUCTION

When the performance of an IR system on a query can be accurately predicted, an informed decision can be made as to whether the query should be expanded, reformulated, biased toward a particular intent or altered in some other way. Increasing evidence points to the fact that valuable clues to a query's ambiguity and quality of corresponding results can be gleaned from query pre-retrieval features, and post-retrieval properties of the query's result set [4]. For example, the *query clarity* score [3] measures the divergence of a language model over the top $\mathcal{D}$ pages from the generic language model of the collection.

However, a significant drawback of methods that analyze the result set is that they must perform an initial retrieval. Additionally, most proposed methods must also process the full text of each document, which adds extra processing time to each query. The search engine must fetch the full text of each top-ranked document, and then perform additional computation that is proportional to the size of the documents. Since performance prediction is only one part of the entire retrieval process, adding computational load at intermediary steps is undesirable, especially in applications like Web search where speed is critical. Thus, a compelling research question is what benefits of result-set analysis like query clarity can be retained with less computational cost.

We thus investigate the benefits of pre-computing a low-dimensional summary of a document, such as a vector of topic class predictions for a standard topic hierarchy, e.g. ODP [7]. We focus on two query performance measures: *query difficulty*, which measures retrieval risk in terms of the average precision (AP) of the top-ranked results; and *expansion risk*, which measures the absolute value of gain or loss in AP from using query expansion. Predicting the latter directly is an interesting problem since whether or not to do expansion may be the end goal. Furthermore, query difficulty and expansion risk are distinct problems and have been shown [1] to be only weakly correlated. Our analysis below suggests that query performance prediction using class-based analogues may offer results comparable to traditional measures that use full document content.

2. RELATED WORK

Existing work on query performance prediction, summarized in Fig. 1, can be seen as calculating various distances between a global background model of the collection, θ_G, a query model using pre-retrieval features, θ_Q, a language model based on the results of the original query, θ_{R1}, and a language model based on the results of the expanded query, θ_{R2}. We focus on the change in pre- and post-retrieval models relative to the global background model, since this is where the majority of effects are observed. This comprises the arcs Δ_{QG}, a type of query specificity; Δ_{QR1}, a measure of query drift; and Δ_{R1G}, the specificity of results, the analogue of traditional clarity.

A variety of other work has examined query classification and use of class labels. Recently [8] quantified query ambiguity using ODP metadata for individual query terms. Class entropy in result sets has been used to identify ambiguous queries [9] and for advertising decisions [2]. In contrast to these studies our focus is on developing class-based analogues for query performance prediction.

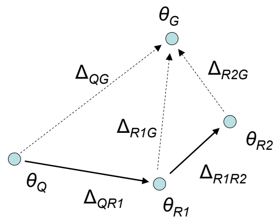
3. METHODS

We use the following two representations to model the language of the query and top-ranked documents.

Unigram language models. This is a $\mathcal{V}$-dimensional vector representing the parameters of a multinomial distribution over the K words in the vocabulary. Typically, no stopping or stemming of words is performed. Model similarity is computed using KL-divergence with Dirichlet smoothing.

Topic prediction vector. A logistic regression classifier trained over a crawl from ODP is used to label every document with 1 to 3 classes depending on whether they sur-

SIGIR '09 Boston, U.S.A.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.



Symbol	Name	Study
Δ_{QG}	Simplified clarity	He & Ounis [5]
Δ_{QR1}	Query drift/coverage	Winaver & Kurland [10]
Δ_{R1G}	Clarity Improved clarity	Cronen-Townsend & Croft [3] Hauff, Murdock, Baeza-Yates [4]
Δ_{R1R2}	Expansion drift	Zhou & Croft[11] (variant)
Δ_{R2G}	Expansion clarity	–

Figure 1: Graphical depiction of model divergences.

passed a threshold (optimized for F1 over validation data). For simplicity, we flattened the top two-levels of the hierarchy and only predict for the $T = 219$ classes that were most frequent in our training set. Model similarity between representations u and v is computed using the metric $\Delta(u, v) = 1/2 \cdot \sum_{i=1}^{T} |u_i - v_i|$. We computed θ_G and θ_{R1} by aggregating the topic distribution for all documents in the collection and result set respectively. We computed θ_Q in two steps. First, we pre-computed topic distributions for each word in the corpus by aggregating the predicted classes of the documents in which the word occurs. Then, for a given query we combined the topic representations for its individual terms using a fuzzy-AND operator.

4. EVALUATION

We used title queries over two TREC Web datasets: wt10g (1.7m pages, topics 451–550) and gov2 (25m pages, topics 701-850). Indexing and retrieval were performed using the Indri system in the Lemur toolkit [6]. To compute the expansion baseline we used the default Relevance model expansion method in Indri 2.2[1], with interpolation parameter $\alpha = 0.5$, feedback with top 50 documents and top 20 expansion terms. We compared how the LM and topic representations affected ability to predict query difficulty and expansion risk, as measured using Kendall's tau correlation with average precision and the absolute magnitude of expansion gain or loss respectively. Results are summarized in Table 1. The LM representation was slightly more effective at predicting query difficulty for both collections, while the TP representation was more effective at predicting expansion risk, especially when topic specificity of a query Δ_{QG} was combined with topic query drift Δ_{QR1}. Fig. 2 shows how these features isolate expansion-neutral queries for wt10g.

5. DISCUSSION & CONCLUSIONS

We explored the use of new sources of evidence in estimating two important measures of query performance, query difficulty and expansion risk, by comparing two document representations – a low-dimensional pre-computed topic representation and a much larger unigram language model – over two standard Web collections. We found that while the LM representation can sometimes give slightly better performance for query difficulty and expansion risk, using pre-computing topic predictions is not far behind. In particular, the topic-based representation is especially effective for pre-retrieval prediction (query classification). This suggests that topic information may often serve as an acceptable, and much more efficient, proxy for predicting query properties and analyzing search results.

	DocRep	Δ_{QG}	Δ_{R1G}	Δ_{QR1}	$\frac{\Delta_{QR1}}{\Delta_{QG}}$	$\frac{\Delta_{R1G}}{\Delta_{QG}}$
wt10g	TP	0.013	0.089	0.077	0.091	0.033
	LM	0.032	0.126	0.256	0.260	$0.161^{\star}$
gov2	TP	0.108^{+}	0.047	0.069	$0.130^{\star}$	0.001
	LM	0.001	$0.137^{\star}$	0.071	0.077	$0.141^{\star}$
wt10g	TP	$0.320^{\star}$	0.052	$0.322^{\star}$	$0.355^{\star}$	$0.330^{\star}$
	LM	$0.250^{\star}$	$0.240^{\star}$	0.071	0.019	$0.225^{\star}$
gov2	TP	0.063	0.124^{+}	0.048	0.070	0.090^{+}
	LM	0.001	0.100^{+}	0.060	0.060	0.100^{+}

Table 1: Kendall-τ correlation of different model similarities with query difficulty (top) and expansion risk (bottom). Document representation (DocRep) is either TP (topic prediction) or LM (language model). Superscripts $\star$ and $+$ denote significance of $p < 0.01$ and $p < 0.10$ respectively.

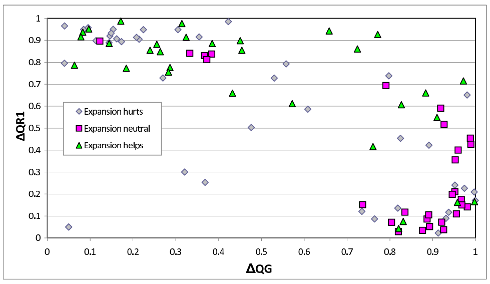


Figure 2: Example showing how expansion-neutral ($< 15\%$ AP gain/loss) wt10g queries (dark squares) typically have high topic specificity (TP:Δ_{QG}) and low post-retrieval topic drift (TP:Δ_{QR1}).

6. REFERENCES

[1] B. Billerbeck. *Efficient Query Expansion.* PhD thesis, RMIT University, Melbourne, Australia, 2005.

[2] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *CIKM '08*, pages 1003–1012.

[3] S. Cronen-Townsend and W. Croft. Quantifying query ambiguity. In *Proceedings of HCL 2002*, pages 94–98, 2002.

[4] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *CIKM '08*, pages 439–448.

[5] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31:585–594, 2006.

[6] Lemur. Lemur toolkit for language modeling & retrieval. http://www.lemurproject.org, 2002.

[7] Netscape Communication Corporation. Open directory project. http://www.dmoz.org.

[8] G. Qiu, K. Liu, J. Bu, C. Chen, and Z. Kang. Quantify query ambiguity using ODP metadata. In *SIGIR '07*, pages 697–698.

[9] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *WWW '07*, pages 1169–1170.

[10] M. Winaver, O. Kurland, and C. Domshlak. Towards robust query expansion: model selection in the language modeling framework. In *SIGIR '07*, pages 729–730.

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

[1] Document models were Dirichlet-smoothed, $\mu = 1000$.