Predicting Query Performance via Classification

Kevyn Collins-Thompson Paul N. Bennett
Microsoft Research
1 Microsoft Way
Redmond, WA USA 98052
{kevynct,paul.n.bennett}@microsoft.com

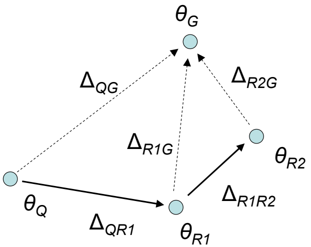
Abstract. We investigate using topic prediction data, as a summary of document content, to compute measures of search result quality. Unlike existing quality measures such as query clarity that require the entire content of the top-ranked results, class-based statistics can be computed efficiently online, 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 two performance-related tasks: predicting 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 information retrieval 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 [9]. For example, the *query clarity* score [7] measures the divergence of a language model over the top-ranked pages from the generic language model of the collection. A separate but related performance prediction problem is to assess the *likely effect of query expansion* for a given query. Because query expansion is both inherently risky and adds further computational expense, methods for predicting the likely success of expansion and correctly scaling back expansion when it is unlikely to be effective are both valuable.

However, existing research in this area has been somewhat incomplete. Figure 1 gives a graphical summary of different pre- and post-retrieval models being compared, highlighting existing and missing work in the current body of research. First, properties of the top-ranked documents retrieved using an expanded query may not only be informative in relation to the original result set but also in relation to pre-retrieval features. Second, while a shift in word distribution between the collection, initial top-ranked results, and expansion results may be informative, because of vocabulary variation these comparisons are necessarily noisy.

To help alleviate this noise in the comparison and capture more of the underlying semantics of queries and documents, we investigate performing a precomputed classification of documents into a set of topics, such as defined by the Open Directory Project (ODP) [15], via models learned from labeled data.



Symbol	Name	Study
$\varDelta_{QG}$	Simplified clarity	He & Ounis [10]
$\varDelta_{QR1}$	Query drift/coverage	Winaver & Kurland [21]
$\varDelta_{R1G}$	Clarity Improved clarity	Cronen-Townsend & Croft [7] Hauff, Murdock, Baeza-Yates [9]
$\varDelta_{R1R2}$	Expansion drift	Zhou & Croft[23] (variant)
$\varDelta_{R2G}$	Expansion clarity	This study.

Figure 1: Graphical depiction of model divergences. The general collection model is shown as θ_G. Inter-model KL-divergences of interest are shown as directed arrows, e.g. $\varDelta_{QG} = KL(Q||G)$, *i.e.*, the distance between the query model and background model. A summary of related work on each model comparison is also shown.

With this pre-computed classification, we later perform fast online comparisons in *topic space* to help restore this focus on semantic distance. Analogous to traditional clarity, we introduce *topic clarity* counterparts to the traditional language model components in Figure 1 and investigate their effectiveness.

A significant drawback of methods that analyze the result set is that they must incur the computational cost of performing an initial retrieval, as well as the cost of processing the full text of each top-ranked document. 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, we also examine whether the benefits of result-set analysis like query clarity can be approximated with less computational cost than using a full language model.

Throughout our analysis, we focus on how effective different model-divergence features are at predicting two types of query performance measure: *query difficulty*, which measures retrieval risk via the average precision (AP) of the top-ranked results; and *expansion risk* which estimates the likely magnitude of the relative gain or loss in AP obtained 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 predicting query difficulty and expansion risk are distinct problems that are only weakly correlated [3].

Our main contributions are as follows. We introduce new models and representations for estimating two important measures of query performance: query difficulty and expansion risk. Our work brings together features from previous studies on query difficulty based on divergences between language models of the query, collection and initial results. We extend this to include a model of *expansion results* from the expanded query. With these models and features, we compare the performance of two model representations: a low-dimensional pre-computed topic representation and a much larger unigram language model over two standard Web collections. We also develop a simple, effective method for deriving a topic representation, modeled as a distribution over ODP categories, of a query by estimating and combining pre-computed topic representations from the individual query terms.

2 Related Work

A number of previous models for query performance prediction can be viewed as special cases within a framework where various distances are calculated 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}. A summary of related work comparing different query and expansion models is shown in Fig. 1.

In Section 3 we give analogues to each of these that can be computed using document topic prediction data. We focus on the change between pre- and post-retrieval models relative to the global background model, since this is where the majority of effects are observed. This comprises the following arcs. The divergence $\varDelta_{QG}$, which we call *simplified clarity*, is a pre-retrieval measure of query specificity and compares the query against the collection model. The post-retrieval divergence $\varDelta_{QR1}$ measures *query drift* in the initial results. The divergence $\varDelta_{R1G}$ is the analogue of traditional clarity, measuring the similarity of the results model to the generic collection model. The divergence $\varDelta_{R2G}$ is a new additional measure that we call *expansion clarity* that estimates the specificity of the expanded results compared to the collection. We also include for completeness $\varDelta_{R1R2}$, the drift from the initial results to the expanded results.

Our examination of model drift extends recent studies that find *variance* to be an important facet of predicting query performance. More specifically, the sensitivity of some aspect of the retrieval process to variation in input or model parameters has been shown to be effective in varying degrees. This includes variance of results ranking (by varying document models) [23], query variation [22], query term *idf* weights [20] and document scores [8]. Aslam & Pavlu [2] introduced variation in the retrieval function instead of the model, by combining TREC runs from multiple systems for the same query.

While the above studies have looked at query difficulty, few have looked at predicting for expansion risk or difficulty. The significant downside risk of query expansion has been noted for decades [17] but has been largely neglected as an evaluation criterion in favor of average performance, with some recent exceptions [6] [14][1]. For query expansion algorithms to become more reliable, it will be important for them to correctly identify and manage risk for queries. We define *expansion risk* in this study to be the magnitude of the relative gain or loss in average precision from applying query expansion, relative to the unexpanded query. Thus, queries with small expansion risk are unlikely to be affected one way or the other by the application of expansion.

A variety of other work has examined query classification and use of class labels. Recently [16] quantified query ambiguity using ODP metadata for individual query terms, and [18] examined the category spread of top-ranked documents to identify ambiguous queries. In contrast to these studies our focus is on establishing and comparing analogues for query performance prediction based on class labels.

3 Methods

Statistics for predicting performance properties of a query can be categorized by the type of observations required to calculate them. Basic *pre-retrieval statistics* use features of the query alone, such as query length or query term *idf* values, without requiring document retrieval using the query [10]. *Post-retrieval statistics* require at least one retrieval step where documents are ranked. The content and/or meta-data of the resulting documents then give us additional information for estimation. The efficiency of post-retrieval statistics depends on the particular document representation used: topic predictions may be pre-computed and do not require fetching or analyzing potentially large documents at run-time. Although document language models may also be precomputed, they use a much larger representation proportional to the vocabulary size. In addition, any document similarity or distance computations for clustering or smoothing are also of correspondingly higher cost.

In the following sections we denote the collection by G, the query by Q, and assume that a set $R1$ of k documents is returned from G in response to Q. Furthermore, after applying query expansion to Q to obtain an expanded query Q', we obtain a set $R2$ of k documents in response to Q'.

Figure 1 shows the models and the relations between them that are of interest in this study. We use the notation $\varDelta_{AB}$ to denote a divergence measure between two models A and B. For example, in the context of language-model based statistics $\varDelta_{AB}$ denotes the KL-divergence $KL(A||B)$ between models A and B. Since KL-divergence is not symmetric, the ordering of A and B is important, and we use an arrow in Figure 1 to specify the direction of comparison.

3.1 Language-Model Based Statistics

As is standard, we use unigram language models as the representation basis for computing the language-model based statistics. This is a K-dimensional vector representing the parameters of a multinomial distribution over the K words in the vocabulary. Model similarity is computed using KL-divergence with Dirichlet smoothing, with KL-divergence defined as $\varDelta(u, v) = \sum_i u_i \log \frac{u_i}{v_i}$ for language model distributions u and v.

3.2 Topic Based Statistics

We chose to use the ODP [15] for classification because of its broad, general-purpose topic coverage and availability of reasonably high-quality training data. Using a crawl of ODP from early 2008, we first split the data into a 70%/30% train/validation set, then identified the topic categories (some categories like "regional" are not topical and were discarded) that had at least 1K documents as good candidates for models that could be learned well and would be broadly applicable – resulting in 219 categories. We leave study of comparing distances in a hierarchy to future work and simply flattened the two levels to a m-of-T (where $T = 219$) prediction task. We then augmented the search index for every document with at least one and up to 3 predictions for each document, assuming the predictions surpass a minimal confidence threshold (approx. 0.05). Thus, minimal index bloat is incurred.

When aggregating the topic distribution for a result set, the topic representation θ is a T-dimensional vector, with one element per ODP class containing the average document class probability for that class. We computed the topic representations θ_G and θ_{R1} by aggregating the topic distribution for all documents in the collection and result set respectively. Model similarity between representations u and v is computed using the 'city block' (or Manhattan) metric

$$\Delta(u, v) = 1/2 \cdot \sum_{i=1}^{T} |u_i - v_i|. \tag{1}$$

We chose this standard symmetric similarity measure due to the nature of the class prediction vector, which unlike language models, is typically not normalized because documents can belong to more than one class, and because magnitude information is important to retain to assess topic prediction confidence.

Because the user's query is expressed in words and not topic categories, we must somehow compute a topic representation θ_Q of the query Q to obtain the pre-retrieval topic-based statistic. We do this in two steps. First, we pre-compute off-line a topic distribution θ_w for each word w in the corpus, by aggregating the predicted classes of the documents in which the word occurs. Then, for a given query we combine the topic representations for its individual terms using an operator of the form

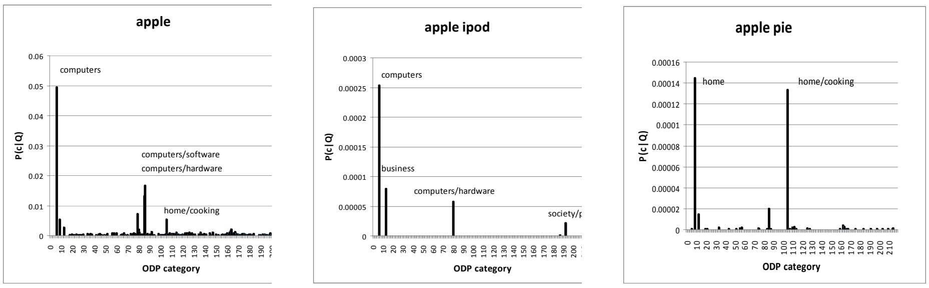
$$\theta_Q[t] \propto \prod_{w \in Q} (\theta_w[t] + \epsilon) \tag{2}$$

which after expanding and collecting like terms in ϵ, can be written as

$$\begin{aligned}\theta_Q[t] \propto \prod_{w \in Q} \theta_w[t] + \epsilon \sum_{w_1 \in Q} \prod_{v \in Q \setminus w_1} \theta_v[t] + \epsilon^2 \sum_{w_1, w_2 \in Q} \prod_{v \in Q \setminus w_1, w_2} \theta_v[t] \qquad (3)\\ + \ldots + \epsilon^{N-1} \sum_{w \in Q} \theta_w[t] + \epsilon^N. \qquad (4)\end{aligned}$$

The parameter ϵ controls the conjunctive behavior of the operator: setting $\epsilon = 0$ gives a pure multiplicative AND operator, and increasing ϵ relaxes this condition, so that $\epsilon = 1$ gives all subsets of Q's terms equal weight. Large values of $\epsilon >> 1$ give increasing OR-like behavior that emphasize the sum over terms, rather than the product. In our experiments we focus on conservative AND-like behavior by using a value of $\epsilon = 0.001$. This approach is easily and efficiently generalized to an inference network, where the query terms are evidence nodes and a richer set of operators is possible, such as those in the Indri retrieval system [19].

Examples of the resulting topic distribution for three different queries are shown in Figure 2. The horizontal axis gives the (flattened) ODP level 1 and 2 categories, while the vertical axis gives $P(c|Q)$, the probability of category c given the query Q. We note that this new pre-retrieval topic-based query representation has many uses beyond performance prediction applications, such as providing an additional set of features for estimating query similarity.



(a) Topic distribution for query ‘apple’

(b) Topic distribution for query ’apple ipod’

(c) Topic distribution for query ’apple pie’

Figure 2: Example showing how the ODP category distribution profiles for different queries can reflect ambiguity or clarity in topic. The ambiguous query ‘apple’ has two main senses of ‘computer’ and ‘home/cooking’, with the computer sense predominating. Refining the ‘apple’ query to ‘apple ipod’ (2b) focuses on the computer topic, while refining to ‘apple pie’ (2c) focuses on the ‘cooking’ sense.

4 Evaluation

Our evaluation is structured as follows. After describing our datasets and experimental setup, we first examine the cases where the topic (TP) representation produces features with comparable predictive power to their language model (LM) counterparts. We do this for both query difficulty and expansion risk prediction tasks. Second, we examine the predictive power of the information in the results from the *expanded* query via the resulting new expansion clarity feature, the divergence Δ_{R2G}, as well as the related expansion drift feature Δ_{R1R2}. Third, we examine combined models in which both TP and LM features are used to predict query difficulty and expansion risk.

4.1 Datasets and experimental setup

Our evaluation is based on two TREC Web datasets that have been widely used for query difficulty prediction: wt10g (1.7m pages, topics 451–550) and gov2 (25m pages, topics 701-850). Also, query performance prediction is known to be more difficult for these Web topics [9]. Indexing and retrieval were performed using the Indri system in the Lemur toolkit [13].

Our queries were derived from the title field of the TREC topics. Phrases were not used. We wrapped the initial query terms with Indri's `#combine` operator, performed Krovetz stemming, and used a stoplist of 419 common English words. To compute the query expansion baseline we used the default expansion method in Indri 2.2, which first selects terms using a log-odds calculation, then assigns final term weights using the Relevance Model [12]: document models were Dirichlet-smoothed with $\mu = 1000$. Indri's feedback model is linearly interpolated with the original query model weighted by a parameter α. By default we used the top 50 documents for feedback and the top 20 expansion terms, with the feedback interpolation parameter $\alpha = 0.5$ unless otherwise stated.

4.2 Comparing topic and language model representations

Our goal in this section is to compare the predictive power of TP and LM representations for the model divergence features shown in Figure 1 as well as some basic pairwise ratios of these features.

Query difficulty The Kendall's tau correlations with average precision for each feature are shown in Table 1. We note that our relatively low query clarity Δ_{R1G} correlation is in line with published studies using similar methods [9] for the same collection. On both collections, the LM version of traditional query clarity Δ_{R1G} gave a higher correlation with AP than its TP counterpart. Performance for the post-expansion drift feature Δ_{R1R2}, however, was not only better than query clarity, but TP and LM performance was comparable: the TP improvement over LM for Δ_{R1R2} was significant for gov2 and statistically equivalent for wt10g. The best performing TP feature on both wt10g and gov2 was Δ_{R1R2} (correlation = 0.11 and 0.25 respectively). The best performing LM feature on wt10g was Δ_{QR1}/Δ_{QG} (correlation = 0.26) and for gov2 $\Delta_{R2G}/\Delta_{R1G}$ (correlation = 0.20).

	DocRep	Δ_{QG}	Δ_{R1G}	Δ_{QR1}	Δ_{R1R2}	Δ_{R2G}	$\frac{\Delta_{QR1}}{\Delta_{QG}}$	$\frac{\Delta_{R1G}}{\Delta_{QG}}$	$\frac{\Delta_{R2G}}{\Delta_{R1G}}$
wt10g	TP	0.013	0.089	0.077	0.110	0.060	0.091	0.033	0.000
	LM	0.032	0.126	0.256	0.140	0.026	0.260*	0.161*	0.231$^+$
gov2	TP	0.108$^+$	0.047	0.069	0.250*	0.010	0.130*	0.001	0.100
	LM	0.001	0.137*	0.071	0.151	0.011	0.077	0.141*	0.204$^+$

Table 1: Query difficulty: Predictive power of different model divergence features according to Kendall-τ correlation with average precision. 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.

Expansion risk Recall that expansion risk is defined as the magnitude of the relative gain or loss in average precision of applying the expansion algorithm, compared to the unexpanded query. Kendall's-tau correlations are shown in Table 2. Although LM-based features were more effective at predicting query

	DocRep	Δ_{QG}	Δ_{R1G}	Δ_{QR1}	Δ_{R1R2}	Δ_{R2G}	$\frac{\Delta_{QR1}}{\Delta_{QG}}$	$\frac{\Delta_{R1G}}{\Delta_{QG}}$	$\frac{\Delta_{R2G}}{\Delta_{R1G}}$
wt10g	TP	0.320*	0.052	0.322*	0.300*	0.169$^+$	0.355*	0.330*	0.280*
	LM	0.250*	0.240*	0.071	0.260*	0.150$^+$	0.019	0.225*	0.110
gov2	TP	0.063	0.124$^+$	0.048	0.260*	0.040	0.070	0.090$^+$	0.188$^+$
	LM	0.001	0.100$^+$	0.060	0.201*	0.040	0.060	0.100$^+$	0.281*

Table 2: Expansion risk: Kendall-τ correlation of different model divergence features. 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.

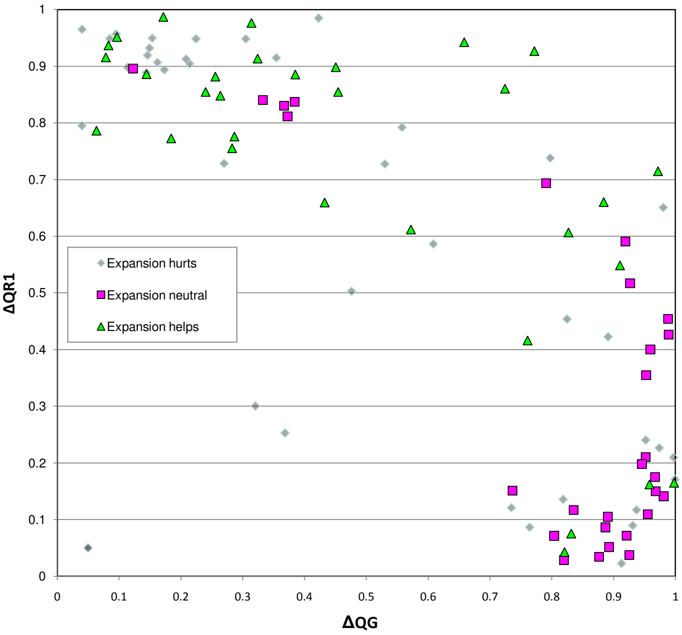
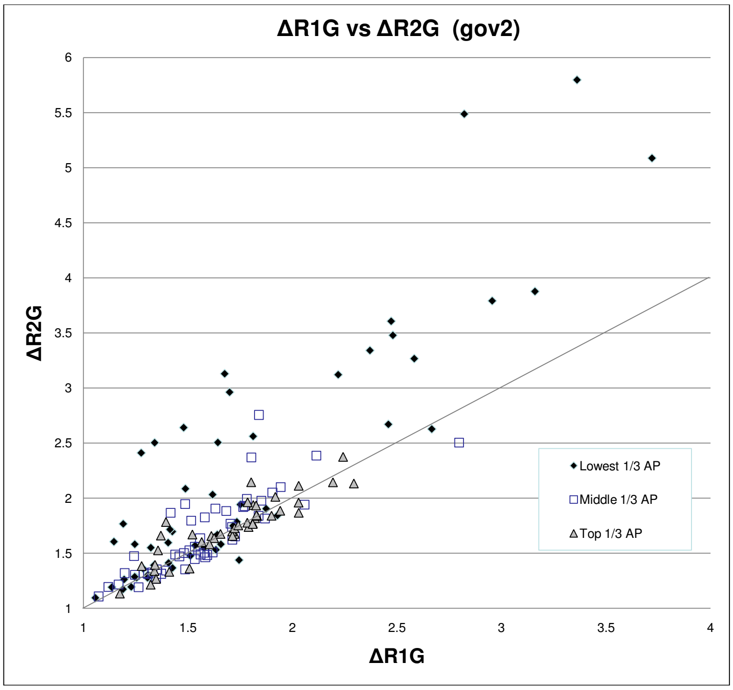


Figure 3: Example showing how expansion-neutral (< 15% AP gain/loss) wt10g queries (dark squares) typically have high topic specificity (TP:Δ_{QG}) and low post-retrieval topic drift (TP:Δ_{QR1}).

difficulty, TP-based features were generally more effective at predicting expansion risk, especially when multiple features were combined (e.g. as in the ratio Δ_{QR1}/Δ_{QG}). Figure 3 shows how combining information from both Δ_{QR1} and Δ_{QG} helps isolate expansion-neutral queries – those queries for which expansion is unlikely to be effective. Queries with higher Δ_{QG} are more specific, being farther from the general collection model. At the same time, queries with low topic query drift Δ_{QR1} have results that match the expected topic profile based on the query terms alone. In this example, queries that are both topic-specific and with focused results are more unlikely to be affected by applying query expansion.

4.3 Predictive power of expansion clarity ($\boldsymbol{\Delta_{R2G}}$) and expansion drift ($\boldsymbol{\Delta_{R1R2}}$) features

The expansion clarity (Δ_{R2G}) and expansion drift (Δ_{R1R2}) features are interesting because they use additional new 'post-expansion' evidence: the results of the expanded query, not just the initial query. We find that such post-expansion features are indeed more effective and stable than features based only on pre-expansion models. For example, the expansion drift feature Δ_{R1R2}, which is dependent on both initial and expansion results models, is remarkably effective and stable across all tasks, representations, and collections compared to any pre-expansion feature. Looking at the Δ_{R1R2} column in Tables 1 and 2, we can see



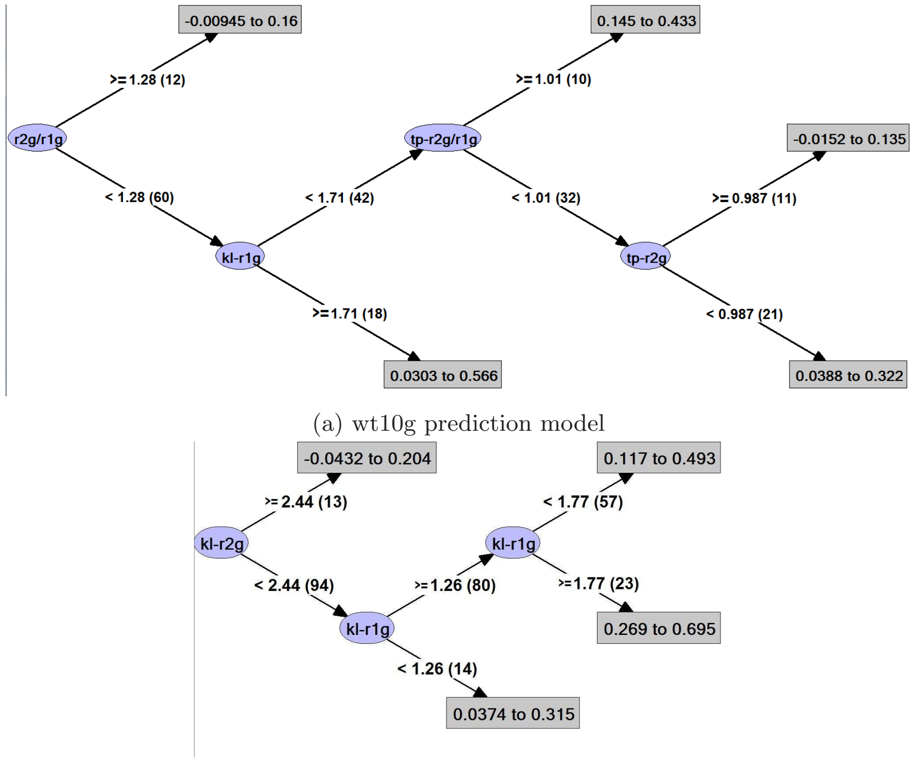
(a) Query clarity ($\varDelta_{R1G}$) vs. expansion clarity ($\varDelta_{R2G}$)

Figure 4: Queries whose initial results clarity ($\varDelta_{R1G}$) is hurt by expansion (higher $\varDelta_{R2G}$) appear as points above the line and are substantially more likely to have poor initial average precision. Query clarity is on the x-axis and expansion clarity on the y-axis. Shown are queries partitioned into the lowest-, mid-, and highest-scoring (AP) third for the gov2 corpus. Results for wt10g are similar and not shown for space reasons.

that the $\varDelta_{R1R2}$ feature is consistently among the best-performing features for either TP or LM representations: it is the top-performing TP feature for predicting both wt10g and gov2 query difficulty and for gov2 expansion risk (with excellent performance on wt10g). For LM, in 3 out of 4 cases it is one of the top two best features, second only to the ratio $\frac{\varDelta_{R2G}}{\varDelta_{R1G}}$, which uses the additional information about the collection model. Figure 4 gives further insight into how adding expansion clarity $\varDelta_{R2G}$ to the basic query clarity $\varDelta_{R1G}$ feature helps discriminate the most difficult queries more effectively than query clarity alone.

4.4 Combining topic- and LM-based features for prediction

To analyze the interaction of the input variables, we used the WinMine v2.5 toolkit [5] to build a regression-based predictor of average precision using a 70/30 train/test split. In particular, we used WinMine to build dependency networks – essentially a decision tree built using a Bayesian machine learning algorithm [4, 11]. We chose this representation for its amenability to qualitative analysis. Note that a more direct comparison to the ranking correlations presented above would



(a) wt10g prediction model

(b) gov2 prediction model

Figure 5: Prediction models for query difficulty for the wt10g (top) and gov2 (bottom) collections, estimated using a Bayesian decision tree learning algorithm, using both topic and LM model divergences as input features. Inequalities at the branches indicate the threshold value for the input variable (shown in the ellipses). The resulting value of the target variable (average precision) is shown with a range of one standard deviation in the shaded rectangle. Both models have selected Δ_{R1G} (query clarity) and Δ_{R2G} (expansion clarity) together as the primary factors in predicting average precision.

require training a ranking model. We defer that problem to future work and simply present the harder task of predicting the actual value of the dependent variable as a means of studying the interaction of the input variables.

The resulting decision trees for query difficulty on the gov2 and wt10g corpora are shown in Figure 5. The measure of accuracy was Root Mean Squared Error (RMSE), where the dependent variable to be predicted was average precision and the input variables were the model divergences from Figure 1 for both the topic and language model representations. Models for both corpora were able to attain better performance (lower RMSE) than a default baseline that simply predicted the distribution mean. The gov2 model using all the features attained an RMSE of 0.147, compared to a default baseline with a higher RMSE of 0.198. The wt10g model using all the features attained an RMSE of 0.163, compared to a default baseline RMSE of 0.175.

While the specific models estimated for each collection are different, they both rely exclusively on the Δ_{R1G} and Δ_{R2G} divergences (or the ratio between them) as the two primary prediction factors, ignoring Δ_{QG}, Δ_{QR1}, and Δ_{R1R2} in both topic and LM representations. This suggests that query clarity and expansion clarity together are most effective at summarizing the trajectory of topic drift that occurs when query expansion is applied, compared to other features, or either clarity feature alone. We also note that the model estimated using wt10g relies on a combination of both topic and language model features to achieve lower RMSE, making use of the complementary aspects of these representations.

5 Discussion and Conclusion

A significant amount of implicit pre-computation lies behind the topic-based representation of a document or query: from the training of the ODP category classifier from thousands of examples, to index-time labeling of topics for individual pages, and aggregating these for building a run-time mapping from terms to topic distributions. A similar effect might be accomplished for the language model representation by learning a global term-term translation matrix to smooth the query model, but with a corresponding increase in size and complexity, moving from a few hundred static ODP categories for the topic representation, to potentially hundreds of thousands of co-occurring terms per word for the language model representation.

Other algorithms for distilling a topic representation of a query are certainly possible: adding phrases, or making more subtle distinctions between morphological variants of terms, becomes important since typical Web queries are usually less than five words long. For example, the topic distributions for 'cat' and 'cats' (independent of other query terms) could be quite different (e.g. since 'cats' is a musical theatre title). This processing would take place at indexing time and thus could make use of large linguistic resources, and potentially increased computation that might not be practical at query run-time. Using latent topic models trained on the corpus could be another possible research direction, in cases where the corpus is of manageable size or effective methods of sampling representative content are available. Such probabilistic models would also have the advantage of giving another principled way to estimate the probability of a topic given a set of terms.

The evaluation shows that while the LM representation can sometimes give slightly better performance for query difficulty, using pre-computing topic predictions is not far behind for some features. In particular, the topic-based representation is more effective for pre-retrieval prediction (query classification) and superior for predicting expansion risk. This suggests that topic information may often serve as an acceptable, and much more efficient, proxy for predicting query properties and analyzing search results. In addition, our analysis also revealed the value of estimating expansion clarity – the divergence between the expansion top-ranked results and the collection – in either representation, with post-expansion features such as expansion drift being highly effective and stable.

Future applications of topic-based representations include more robust, efficient query similarity measures and measures of result diversity. Also interesting to consider are more sophisticated inference methods for estimating a topic distribution from a query based on the use of additional term dependency features.

Acknowledgments

We thank Sue Dumais for valuable discussions and feedback on this work.

References

1. G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR '04*, pages 127–137.
2. J. A. Aslam and V. Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of ECIR '07*, pages 198–209.
3. B. Billerbeck. *Efficient Query Expansion*. PhD thesis, RMIT University, Melbourne, Australia, 2005.
4. D. Chickering, D. Heckerman, and C. Meek. A Bayesian approach to learning Bayesian networks with local structure. In *UAI '97*, pages 80–89.
5. D. M. Chickering. The winmine toolkit. Technical Report MSR-TR-2002-103, Microsoft, Redmond, WA, 2002.
6. K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of SIGIR '07*, pages 303–310.
7. S. Cronen-Townsend and W. Croft. Quantifying query ambiguity. In *Proceedings of HCL 2002*, pages 94–98.
8. F. Diaz. Performance prediction using spatial autocorrelation. In *Proceedings of SIGIR '07*, pages 583–590.
9. C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proceedings of CIKM '08*, pages 439–448.
10. B. He and I. Ounis. Query performance prediction. *Information Systems*, 31:585–594, 2006.
11. D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.
12. V. Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts, Amherst, 2004.
13. Lemur. Lemur toolkit for language modeling & retrieval. http://www.lemurproject.org, 2002.
14. D. Metzler and W. B. Croft. Latent concept expansion using Markov Random Fields. In *Proceedings of SIGIR '07*, pages 311–318.
15. Netscape Communication Corp. Open directory project. http://www.dmoz.org.
16. G. Qiu, K. Liu, J. Bu, C. Chen, and Z. Kang. Quantify query ambiguity using ODP metadata. In *Proceedings of SIGIR '07*, pages 697–698.
17. A. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.
18. R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *Proceedings of WWW '07*, pages 1169–1170.
19. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.
20. V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. On ranking the effectiveness of searches. In *Proceedings of SIGIR '05*, pages 398–404.
21. M. Winaver, O. Kurland, and C. Domshlak. Towards robust query expansion: model selection in the language modeling framework. In *Proceedings of SIGIR '07*, pages 729–730.
22. E. YomTov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty. In *Proceedings of SIGIR '05*, pages 512–519.
23. Y. Zhou and W. B. Croft. Ranking robustness: a novel framework to predict query performance. In *Proceedings of CIKM '06*, pages 567–574.