Garrick Sherman

ABSTRACT

1. INTRODUCTION

Query performance prediction (QPP) is the task of estimating the success of an information retrieval (IR) system in retrieving results for a submitted query. Prediction strategies generally use properties of the query or of the retrieved document set to quantify the success of a system. The former is called pre-retrieval prediction, while the latter is called post-retrieval prediction. On the whole, post-retrieval prediction outperforms pre-retrieval prediction, and our predictor belongs to this class.

In general, post-retrieval prediction methods tend to associate some property of the result set with the relevance of the result set in general (or, more accurately, some measurement thereof). These properties are usually intuitively suggestive of some form of stability or robustness, as discussed in Section 2.

In this work, we propose a predictor inspired by pseudo-relevance feedback techniques that directly incorporates relevance metrics in the prediction process. We accomplish this by asserting that the top r retreived documents in a ranked list for a query Q are relevant to that query; we then use these pseudo-relevance judgments to estimate document retrieval evaluation using natural variations on the query: the top n ranked documents.

2. RELATED WORK

A great deal of research has focused on QPP. The most famous predictor is query clarity, which is simply the KL-divergence of an expanded query model and the collection as a whole [**?**]. We incorporate query clarity as a baseline in our work.

One class of query performance predictor that is related to our work is query perturbation [**?**, **?**, **?**]. These methods alter the query and measure the similarity of the results list across query variations; if the result list remains relatively constant, the system is believed to have little difficulty with the query. Our proposed method is fundamentally a form of query perturbation, with the alternative query forms derived from feedback documents.

3. PSEUDO-RELEVANCE PERFORMANCE PREDICTOR

Our query performance predictor, which we call the pseudo-relevance performance predictor (PRPP), is inspired by the pseudo-relevance feedback technique of treating the top ranked documents as relevant to the query. Doing so allows us to directly model the robustness of query performance metrics to query perturbation. The broad steps of our approach are as follows:

1. Issue query Q against collection C_p; call the results list R_p
2. Record the top r documents in R_p as relevant using each's normalized retrieval score as the degree of relevance
3. Issue Q against target collection C_t, if $C_p \neq C_t$; call the results list R_t
4. For each document D in R_t,
   (a) Construct a pseudo-query Q_D consisting of the top q non-stop terms[1] in D
   (b) Issue Q_D against C_p; call the results list R_{pD}
   (c) Score R_{pD} using the pseudo-relevance judgments from Step 2.
5. Average the scores to compute a single score for Q

Several points in the preceding would benefit from clarification.

First, our model makes no assertions about the identity of the performance collection C_p. While it is valid to use the target collection as the performance collection, we hypothesize that the use of an external collection may yield better results. We investigate this idea in Section **??**.

Next, both r and q are free parameters in our approach. In fact, there are two other parameters in our approach: k controls the length of R_t, while n controls the length of R_{pD}. In general, all parameters should be set empirically. Although not required mathematically, we only allow values of n where $n \leq r$ since setting n greater than r prevents us from retrieving all relevant documents.

One advantage to our model is the ability to predict specific retrieval metrics by estimating levels of relevance. Because these levels are normalized log probabilities, they consist of continuous values (in contrast to TREC qrels which

[1]We remove terms in the standard Indri stoplist: http://www.lemurproject.org/stopwords/stoplist.dft

Corpus	Pred.	MAP		nDCG@20	
		Med.	Max	Med.	Max
AP	QC	**0.5280**	**0.6148**	0.3564	0.4284
	QF	0.4240	0.5637	0.3202	0.4467
	PRPP	0.5030	0.6115	**0.4025**	**0.4880**
wt10g	QC	0.2232	0.3636	0.0943	0.3436
	QF	0.1624	0.3847	0.0881	0.3613
	PRPP	**0.2850**	**0.4615**	**0.2529**	**0.4173**
Robust	QC	0.4240	0.5114	0.3129	0.3914
	QF	0.2203	0.5073	0.1708	0.4926
	PRPP	**0.4381**	**0.5791**	**0.3222**	**0.5047**

Table 1: Median and maximum Pearson's r correlation of predictors for each collection. Bolded values are the largest observed per corpus/metric summary.

Corpus	Pred.	MAP		nDCG@20	
		Med.	Max	Med.	Max
AP	QC	0.3609	0.4440	0.2166	0.2807
	QF	0.3259	0.4215	0.2042	0.3100
	PRPP	**0.4132**	**0.4962**	**0.2701**	**0.3378**
wt10g	QC	0.1561	0.3093	0.0712	0.2622
	QF	0.1599	0.2843	0.0657	0.2411
	PRPP	**0.2710**	**0.3642**	**0.1935**	**0.3053**
Robust	QC	0.2985	0.3714	0.2266	0.2889
	QF	0.1895	0.3745	0.1389	**0.3646**
	PRPP	**0.3433**	**0.4273**	**0.2478**	0.3634

Table 2: Median and maximum Kendall's τ correlation of predictors for each correlation. Bolded values are the largest observed per corpus/metric summary.

contain binary or discrete relevance grades), but are still valid for metrics that incorporate graded relevance judgments like normalized discounted cumulative gain (nDCG). Metrics that use only binary relevance metrics, like mean average precision (MAP) follow the typical procedure of treating all judgments greater than zero as relevant.

4. EVALUATION

4.1 Data

We evaluate our procedure using a variety of TREC datasets that represent a variety of corpora properties:

- AP newsire with topics 101-200 from TREC disks 1 and 2
- Robust with the 2004 topics from TREC disks 4 and 5
- wt10g with the topics 451-500 from the 2000 and 2001 TREC Web tracks

We also make use of Wikipedia[2] as an external collection in some experiments.

5. RESULTS

6. CONCLUSIONS

[2]http://en.wikipedia.org

7. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant No. [blind]. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.