A Survey of Pre-Retrieval Query Performance Predictors

Claudia Hauff
Human Media Interaction
University of Twente
c.hauff@ewi.utwente.nl

Djoerd Hiemstra
Database Group
University of Twente
hiemstra@cs.utwente.nl

Franciska de Jong
Human Media Interaction
University of Twente
f.m.g.dejong@utwente.nl

ABSTRACT

The focus of research on query performance prediction is to predict the effectiveness of a query given a search system and a collection of documents. If the performance of queries can be estimated in advance of, or during the retrieval stage, specific measures can be taken to improve the overall performance of the system. In particular, pre-retrieval predictors predict the query performance before the retrieval step and are thus independent of the ranked list of results; such predictors base their predictions solely on query terms, the collection statistics and possibly external sources such as WordNet. In this poster, 22 pre-retrieval predictors are categorized and assessed on three different TREC test collections.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Performance

Keywords

query performance prediction

1. INTRODUCTION

Research on query performance prediction deals with the task of predicting how well or poorly a query will perform given a search system and a collection of documents. If the performance of queries can be estimated in advance of or during the retrieval stage, specific steps can be taken to improve retrieval.

Over the last years, a number of query performance prediction algorithms have been proposed that broadly fall into two categories: pre-retrieval and post-retrieval approaches. Pre-retrieval predictors estimate the performance of a query before the retrieval stage is reached and are, thus, independent of the ranked list of results; essentially, they are search-independent. They base their predictions on query term characteristics, collection statistics and possibly external sources such as WordNet[1], which provides information on the terms' semantic relationships. Since pre-retrieval predictors rely on information that is available at indexing time, they usually can be calculated more efficiently, causing less overhead to the search system. On the other hand, post-retrieval predictors such as Query Clarity[3] base their predictions on the ranked list of results, which provides more information to the predictor, making accurate predictions easier to achieve. This benefit is offset however, by the added requirement of a second retrieval stage, as, given unsatisfactory predictor scores, another retrieval stage might be necessary. In this poster, we concentrate on the evaluation of existing pre-retrieval predictors. In the following section, a categorization of pre-retrieval predictors is presented. Subsequently, an overview of the predictors' performances across different test collections is given, followed by a brief discussion of the results.

2. CATEGORIZATION

Pre-retrieval predictors can be divided into four different groups according to the heuristic they exploit: *specificity*, *ambiguity* (AMBI), *term relatedness* (REL) and *ranking sensitivity* (RNK). The specificity based predictors predict a query to perform better with increased specificity. How the specificity is determined, further divides these predictors into collection statistics based and query based. The former are based on the inverse term/document frequency of the query terms (e.g. $AvIDF$, $AvICTF$, SCS) while the predictor in the latter case ($AvQL$) relies on the length of the query terms alone. A different variety of predictors exploits the query terms' *ambiguity*. In such a scheme, if a term always appears in the same or similar contexts, the term is considered to be unambiguous. Low ambiguity indicates an easy query. $AvQC$ and $AvQCG$ for instance depend on document clustering to determine a term's ambiguity. An alternative to collection based ambiguity is the utilization of WordNet[1]. AvP exploits the number of noun senses in WordNet for prediction, whereas AvP relies on the number of senses across all word types. Here, the higher the sense count, the higher the term's ambiguity. *Term relatedness* based predictors consider the relationship between query terms. A strong relationship between query terms suggests a well performing query. The maximum and average of the pointwise mutual information over all query term pairs ($MaxPMI$, $AvPMI$) are two collection based predictors in this category. The semantic similarity between the query terms (their relatedness) can also be derived from WordNet. The predictors $AvLesk$, $AvVP$ and $AvPath$ belong to this group. Finally, *ranking sensitivity* based predictors

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

predict a query to be difficult if the retrieval algorithm cannot distinguish the documents containing the query terms from each other. Specifically, the predictors in [11] ($AvVAR$, $MaxVAR$, $SumVAR$) rely on the distribution of TF.IDF weights. Note that the last three predictors and the predictors based on document clustering ($AvQC$, $AvQCG$) require a significant amount of preprocessing, compared to predictors exploiting the inverse term/document frequencies alone.

3. EXPERIMENTS

22 pre-retrieval predictors were assessed on three different TREC collections: TREC Volumes 4+5 (minus CR) with title topics 301-450, WT10g with title topics 451-550 and GOV2 with title topics 701-850. The collections were Krovetz stemmed and stopwords were removed. As retrieval approach, Language Modeling with Dirichlet smoothing [10] was chosen. Three different levels of smoothing were selected: low ($\mu = 100$), medium ($\mu = 1500$) and high ($\mu = 5000$). As most predictors exploit the inverse term/document frequencies in some way, it is hypothesized that the amount of smoothing influences the quality of the predictors. For the purpose of evaluating the predictors, we report the linear correlation coefficient (Table 1) and Kendall's tau (Table 2). Results marked with a star are statistically significant; the best performing predictor for each collection and level of smoothing is shown in italics.

		TREC Vol. 4+5			WT10g			GOV2		
		μ100	μ1500	μ5000	μ100	μ1500	μ5000	μ100	μ1500	μ5000
SPECIFICITY	AvQL[6]	0.13	0.14	0.16	-0.11	-0.14	-0.12	-0.05	0.02	0.03
	AvIDF[3]	_0.52*_	0.53*	0.59*	0.21*	0.18	0.18	0.37*	0.32*	0.39*
	MaxIDF[9]	_0.52*_	_0.54*_	_0.60*_	0.31*	0.30*	0.30*	0.35*	0.35*	0.43*
	DevIDF[4]	0.22*	0.24*	0.26*	0.21*	0.25*	0.27*	0.14	0.20*	0.27*
	AvICTF[4]	0.50*	0.50*	0.56*	0.20	0.16	0.16	0.34*	0.30*	0.37*
	SCS[4]	0.49*	0.49*	0.55*	0.15	0.13	0.13	0.31*	0.26*	0.34*
	QS[4]	0.42*	0.42*	0.47*	0.09	0.05	0.05	0.26*	0.18*	0.22*
	AvSCQ[11]	0.25*	0.27*	0.31*	0.32*	0.30*	0.30*	0.40*	0.36*	0.39*
	SumSCQ[11]	-0.01	0.00	0.00	0.20*	0.18	0.15	0.23*	0.23*	0.19*
	MaxSCQ[11]	0.32*	0.35*	0.38*	_0.36*_	0.41*	0.45*	0.39*	0.42*	_0.46*_
AMBI	AvQC[5]	0.45*	0.47*	0.51*	0.18	0.17	0.17	0.28*	0.31*	0.38*
	AvQCG[5]	0.33*	0.34*	0.37*	0.00	-0.03	-0.03	0.04	0.05	0.08
	AvNP[6]	-0.20*	-0.23*	-0.26*	-0.09	-0.10	-0.10	-0.06	-0.04	-0.05
	AvP	-0.11	-0.12	-0.14	-0.17	-0.18	-0.17	0.02	0.01	0.00
REL	AvPMI	0.37*	0.35*	0.39*	0.33*	0.28*	0.26*	0.26*	0.29*	0.33*
	MaxPMI	0.30*	0.30*	0.33*	0.31*	0.27*	0.24*	0.28*	0.31*	0.32*
	AvLesk[2]	0.24*	0.25*	0.27*	0.00	0.01	0.02	0.04	0.08	0.11
	AvPath[8]	0.12	0.14	0.16	0.01	0.04	0.05	-0.02	0.03	0.07
	AvVP[7]	0.25*	0.25*	0.27*	-0.06	-0.06	-0.05	-0.01	0.09	0.13
RNK	AvVAR[11]	0.50*	0.52*	0.56*	0.29*	0.29*	0.30*	_0.43*_	0.40*	0.42*
	SumVAR[11]	0.28*	0.30*	0.31*	0.31*	0.29*	0.28*	0.33*	0.34*	0.30*
	MaxVAR[11]	0.48*	0.52*	0.54*	_0.36*_	_0.42*_	_0.47*_	0.40*	_0.43*_	_0.46*_

Table 1: Results of the predictor evaluations given by the linear correlation coefficient.

		TREC Vol. 4+5			WT10g			GOV2		
		μ100	μ1500	μ5000	μ100	μ1500	μ5000	μ100	μ1500	μ5000
SPECIFICITY	AvQL[6]	0.05	0.04	0.06	-0.01	-0.01	0.00	-0.03	0.04	0.04
	AvIDF[3]	0.30*	0.30*	0.35*	0.22*	0.23*	0.23*	0.27*	0.24*	0.28*
	MaxIDF[9]	0.34*	0.35*	0.40*	0.25*	0.28*	0.30*	0.26*	0.25*	0.32*
	DevIDF[4]	0.17*	0.18*	0.20*	0.13	0.16*	0.18*	0.11	0.15*	0.20*
	AvICTF[4]	0.27*	0.27*	0.32*	0.20*	0.20*	0.22*	0.25*	0.22*	0.27*
	SCS[4]	0.26*	0.26*	0.31*	0.18*	0.17*	0.18*	0.22*	0.20*	0.25*
	QS[4]	0.20*	0.20*	0.23*	0.12	0.08	0.09	0.19*	0.14*	0.16*
	AvSCQ[11]	0.21*	0.21*	0.25*	0.24*	0.24*	0.24*	0.27*	0.24*	0.26*
	SumSCQ[11]	0.07	0.08	0.09	0.14*	0.13*	0.12	0.18*	0.17*	0.14*
	MaxSCQ[11]	0.33*	0.34*	0.38*	0.28*	_0.34*_	_0.37*_	0.27*	0.28*	0.33*
AMBI	AvQC[5]	0.30*	0.30*	0.34*	0.20*	0.22*	0.24*	0.27*	_0.29*_	_0.34*_
	AvQCG[5]	0.31*	0.31*	0.35*	0.18*	0.21*	0.22*	0.27*	0.28*	0.33*
	AvNP[6]	-0.14*	-0.17*	-0.20*	-0.06	-0.05	-0.05	-0.02	0.00	0.00
	AvP	-0.10	-0.12*	-0.15*	-0.11	-0.12	-0.12	0.00	0.02	0.01
REL	AvPMI	0.22*	0.23*	0.27*	0.24*	0.20*	0.19*	0.19*	0.22*	0.24*
	MaxPMI	0.22*	0.22*	0.25*	0.24*	0.23*	0.21*	0.21*	0.23*	0.22*
	AvLesk[2]	0.11	0.11	0.10	-0.01	-0.03	-0.02	0.02	0.04	0.03
	AvPath[8]	-0.04	-0.03	-0.02	0.03	0.03	0.05	-0.09	-0.05	-0.03
	AvVP[7]	-0.03	-0.03	-0.02	-0.04	-0.04	-0.01	0.02	0.05	0.03
RNK	AvVAR[11]	0.36*	0.36*	0.39*	0.25*	0.26*	0.28*	_0.29*_	0.28*	0.29*
	SumVAR[11]	0.28*	0.28*	0.30*	0.21*	0.21*	0.20*	0.22*	0.24*	0.21*
	MaxVAR[11]	_0.40*_	_0.41*_	_0.44*_	_0.29*_	0.33*	0.36*	0.27*	_0.29*_	0.30*

Table 2: Results of the predictor evaluations given by Kendall's tau.

4. DISCUSSION

The predictor performances depend on the particular test collection as well as the particular retrieval approach. Increasing the amount of smoothing tends to increase the predictor performance of those collection based predictors that already perform well on low smoothing. The WordNet based measures are somewhat effective on TREC Volumes 4+5, but fail completely on WT10g and GOV2. Overall, prediction for the title topics of WT10g has shown to be the most difficult. Among the assessed predictors, there is not a single predictor that outperforms all others across all settings evaluated. Although $MaxVAR$ achieves the highest linear correlation coefficient and Kendall's tau for a number of settings, the *difference in correlation* with respect to $MaxIDF$, a predictor that results in slightly lower correlation coefficients, is not statistically significant. This result suggests, and will be studied in more depth in future work, that the comparison of query performance predictors according to the absolute values of their correlation coefficients alone may not be sufficient when the coefficients are similar.

5. REFERENCES

[1] *WordNet - An Electronic Lexical Database*. The MIT Press, 1998.

[2] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI'03*, pages 805–810, 2003.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02*, pages 299–306, 2002.

[4] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE'04*, pages 43–54, 2004.

[5] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *ECIR'08*, pages 689–694, 2008.

[6] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty - a case study on previous trec campaigns. In *ACM SIGIR'05 Query Prediction Workshop*, 2005.

[7] S. Patwardhan and T. Pedersen. Using wordnet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, 2006.

[8] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

[9] F. Scholer, H. Williams, and A. Turpin. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology*, 55(7):637–650, 2004.

[10] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.

[11] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR'08*, pages 52–64, 2008.