Query Performance Prediction Based on Ranking List Dispersion

Joaquín Pérez-Iglesias
Calle Juan del Rosal, 16. 28040 Madrid. Spain.
Universidad Nacional de Educación a Distancia.
<http://nlp.uned.es/jperez>
joaquin.perez@lsi.uned.es

Abstract

In this paper we introduce a novel approach for query performance prediction based on ranking list scores dispersion. Starting from the premise that different score distributions appear for good and poor performance queries, we introduce a set of measures that capture these differences between both types of distributions. The proposed measures will employ the ranking list, output of a search system, as an information source to predict query performance in terms of MAP. The obtained results reveal a significant correlation degree with MAP and are very similar to those achieved with more complex methods. Finally some generic open questions that could guide further research on query prediction methods are introduced.

Keywords: Query Performance Prediction

1. INTRODUCTION

During the last years a growing attention has been focused on the problem of query performance prediction. This topic has turned into an important challenge for the IR community. Query performance prediction deals with the problem of detecting those queries for which a search system would be able to return a document set useful for an user. In other words query prediction objective is the development of a search system able to estimate the quality of its answer before the relevant document set is supplied to the user.

A wide range of possible applications appears for a system like the described before. For example a system could ask the user for some extra information in order to improve the result quality before an answer is supplied; a federated search system could select the best answer from any of its sources; a specialised search system on one specific subject could decide by itself if it needs make use of a broader topic index in order to supply a better answer.

In this paper a novel approach for query performance prediction is introduced. The proposed method falls into post-retrieval prediction methods. This type of predictors make use of the information supplied from the search system, once the search has been carried out, opposite to pre-retrieval prediction where the estimation is computed before the search has been completed. The proposed method is based on the study of document scores distribution among the document scores ranking list. This proposal is based on the hypothesis that some differences between the scores distribution of good performance and poor performance queries can be observed. Some measures that try to capture the prior differences among document scores from a ranking list will be proposed in order to predict query performance.

1.1. Related work

In the last years several works dealing with query performance prediction have been proposed. In general the different prediction methods can be classified into two main groups, those approaches that use information from the results obtained after a query is executed (post-retrieval predictors), and those that try to estimate query difficult before the ranking list is obtained from the search

engine (pre-retrieval predictors). In general it is accepted that the last has as main advantage a low computational cost even at the expense of providing less accurate estimations.

Pre-retrieval predictors use statistics as collection frequency (CF), inverse document frequency (IDF) or query length. These methods try to detect the ambiguity of the query based on these statistics. He and Ounis [9] propose different measures based on IDF and CF as the inverse collection term frequency mean or the IDF standard deviation. Recently Zhao et al. [17] have proposed some measures based on the standard deviation of the query terms, which are weighted using a TF-IDF schema.

Post-retrieval methods are more related to the approach introduced in this paper. First attempts were started by [5] where *Clarity Score* was proposed. This estimator tries to measure the ambiguity of a query with respect to the document collection. The ambiguity of a topic is calculated with Kullback-Leibler divergence (KLD) between the collection and the top ranked documents language model. A good performance query will show a high divergence value, it can be explained by the fact that top ranked documents are about a single topic and this involves low ambiguity. Further works based on Clarity Score like *Ranked List Clarity Score*, which replace ranking scores by ranking position, and *Weighted Clarity Score* that assigns different weights to query terms in order to calculate KLD, appears in [6].

More specific methods for the web environment can be found in [18], *Weighted Information Gain* and *Query Feedback*, where both measures show a good performance for 'ad-hoc' and Named Pages tasks. The first method tries to measure the information gain between a state where an average document is retrieved and the state where the real search has been accomplished. On the other hand *Query Feedback* models the retrieval system as a noisy channel. The input for the noisy channel is the query and the output is the ranking list obtained. From this premise the authors try to measure the degree of noise introduced by the retrieval system and from it to estimate the query performance.

Carmel et al. [4], try to model the relation among the three components that take part into the process of retrieval: topic, set of relevant documents and the collection. The relation between them is measured by means of the Jensen-Shannon divergence. Once the previous measures have been calculated they propose the application of a machine learning method to combine them.

In the works developed by Yom-Tov et. al [16] the proposed model was based on the agreement between the full query and sub-queries, where each one of these sub-queries include only one term from the original query. 'Agreement' is measured from the overlap between full query and sub-queries ranking list results. They conclude that in general hard queries will not show agreement between the obtained ranking list, while a high level of overlapping will mean a good performance query.

Aslam and Pavlu [3] propose a technique based on the Jensen-Shannon divergence between the ranking lists obtained from multiple retrieval functions. This approach is based on the idea that for 'easy' queries ranking functions must agree and therefore a lower divergence would be calculated.

A different approach based on KLD was proposed by Amati et al. [1]. Here the term frequency divergence between top retrieved documents and the whole collection is measured. They claim that a well-defined query (good performance topic), will show a significant divergence.

Recently a new improved version of clarity score has been presented by Hauff et. al [8]. The authors propose two main contributions to clarity score. First, the set of the feedback documents used in order to compute the query language model is fixed to the documents that contain all query terms. Next, they propose to select a subset of terms from the retrieved documents in order to remove the noise generated by those terms with a high document frequency.

Figure 1: 5 Best Performing Topics (left) Vs 5 Worst Performing Topics (right), from Robust 2004 using BM25. Scores have been normalised in $[0, 1]$. The maximum number of retrieved documents has been fixed to 1000.

Following, a set of new post-retrieval predictors based on the scores dispersion among a ranking list will be introduced. These measures will employ the ranking list obtained from a retrieval system in order to predict the query performance, with a low computational cost.

2. RANKING LIST SCORES DISPERSION AS A PREDICTOR

The approach proposed on this paper is based on the study of the ranking list obtained after a retrieval process is executed. As it is well known a ranking function tries to order documents based on their relevance for a topic. For this purpose a ranking function assigns a weight (or score) to each document in the collection. In a naive sense this score can be interpreted as a 'quantitative measure' of the document relevance. The ranking list scores distribution can be an indicative of the quality performance for a specific topic. Based on this premise some differences between document scores distribution for good and poor performing topics should be observed.

For example, if a ranking list has a high value of dispersion among the document scores, it could be a sign that the ranking function has been able to discriminate between relevant and not relevant documents. On the other hand if a low level of dispersion appears, because the ranking function has assigned similar weights, it can be interpreted as it was not able to distinguish between relevant and not relevant documents.

The differences in terms of scores dispersion can be observed in figure 1, where some of the best and worst performance topics for Robust 2004 are represented. As it can be seen best topics show a longer distance between maximum and minimum score and a sharp slope. Opposite to this, topics with poor performance show a lower distance between maximum and minimum and a softer slope.

A feature of this approach is that it can be applied independently of the ranking model employed for retrieval. The reason for this characteristic is due to retrieval models try to maximise in terms of score, the differences between relevant and not relevant documents¹. Based on this property the proposed approach can be considered as a generic method of quality performance prediction, not dependent on the model applied for document weighting.

The measures that will be introduced on next section are focused on trying to capture the differences between good and poor performance topics in terms of dispersion. In order to evaluate the quality of the proposed measures the correlation between them and AP (*average precision*) will be computed.

¹At least for probability based ranking models.

Figure 2: Scores histogram and scores standard deviation respectively for Topic 313. Scores have been normalised in range $[0, 1]$. The maximum number of retrieved documents has been fixed to 1000.

2.1. Proposed Measures

A reliable method to capture and measure dispersion along the obtained ranking list is a key part of this work. Some prior studies have tried to model how document weights are distributed along a ranking list. In general, and as a simplification, it can be assumed that an adequate model could be a mix between an exponential and a normal probability distribution. Exponential for not relevant documents, and normal for relevant documents [12, 13]. Generally a great number of retrieved documents are not relevant (exponential distribution), thus it is likely that a great number of documents will be weighted with a low score. A consequence of it is that ranking lists shape use to hold a long *tail* where a majority of not relevant documents are placed, see figure 1. It is important to understand how the scores distribution will affect some of the typical measures in order to capture the dispersion.

Some notation is needed to define the following proposed measures: (i) A ranking list RL is a document list sorted in decreasing order by their documents scores; (ii) The score assigned to a document d , placed at position i into the ranking list, by a ranking function is defined as $score(d_i)$.

1. **Minimum normalised score:** A first approach to measure the dispersion can be computed with the minor document score found into the ranking list. The minimum score must be normalised in order to make possible a comparison among it and the minimum scores obtained by the rest of topics. Thus for a ranking list of size N :

$$Min = \frac{score(d_N)}{score(d_1)}$$

2. **Standard Deviation:** Standard deviation σ is a simple measure of the variability or dispersion of a data set. A low standard deviation indicates that the data points tend to be very close to the same value (the mean μ), while high standard deviation indicates that the data are spread out over a large range of values. Given ranking list scores mean $\mu(RL)$, standard deviation is computed as next:

$$\sigma(RL) = \sqrt{\frac{1}{N} \sum_{i=1}^N (score(d_i) - \mu(RL))^2}$$

A drawback in the use of the standard deviation is caused by the great number of low scores assigned by the ranking function. As was described previously, a high percentage of document scores have a low value, which causes that mean is displaced towards the region of densest distribution, that is the tail of the ranking list as can be seen in figure 2 (left). As a consequence of it, the deviation on the top documents is not captured properly when the

standard deviation is computed along the full ranking list.

3. **Maximum Standard Deviation:** A different approach to minimise the effect of low scores high frequency is computing the maximum standard deviation. This estimator is based on the idea of computing the standard deviation at each point in the ranking list, and selecting the maximum standard deviation found. As can be seen in figure 2 (right), which shows how standard deviation evolves along ranking list, this measure tends to decrease once the maximum has been reached, coinciding with the start of the *ranking list tail*. The σ_{max} is defined as next:

$$\sigma_{max} = \max[\forall d \in RL, \sigma(RL_{[1,d]})]$$

Following, the experimental setup designed in order to test the validity of the proposed measures will be described.

3. EXPERIMENTAL SETUP

The validity of a query performance predictor is tested based on the correlation coefficient between the proposed predictor and the performance of the real search system in terms of AP. Besides correlation coefficients, a standard test collection and a set of state-of-art retrieval models have been used.

Correlation: On related literature three different correlations coefficients can be found: *Pearson*, *Spearman*(ρ) and *Kendall*(τ). *Pearson* indicates the strength and direction of a linear relationship between two data series. *Kendall* and *Spearman* are based on ranking correlation between both data series. In both cases values are ranked and the correlation coefficient depends on the observable differences between both ranks. More specifically *Spearman* applies a basic linear correlation between both rankings while *Kendall* computes the correlation value counting pairwise swapping needed in order to transform one ranking into the other. The three correlation coefficients calculate a real number in the range $[-1, 1]$, where 1 means perfect correlation, -1 means a perfect inverse correlation and 0 means no correlation at all.

Test data: The different measures proposed in this paper have been tested with the set of documents from TREC Disk4 & 5, minus Congressional Record. This data was employed for the Robust Track 2004 and contains around 528,000 documents with a total size of almost 2 GB [14]. The set of topics are those available from the same track, that is topics 301-450 and 601-700², this makes a total of 249 topics with their relevant judgement. Only the field title from topics has been employed in the experiments.

Ranking models: Since the proposed method can be applied to any retrieval model, we have selected a set of retrieval models representative enough to test the validity and compare the obtained predictions values among them. For each retrieval model³ the parameters have been fixed to the values recommended by Terrier documentation:

- Okapi BM25 [11] with parameters $b = 0.34$, $k_1 = 1.2$ and $k_3 = 8$.
- Language Model proposed by Hiemstra in [10] with $\lambda = 0.15$.
- DFR-PL2 proposed by Amati et. al in [2] with $c = 9.150$.

4. RESULTS

The obtained results⁴ after the execution of the experiments are shown at table 1. These measures were computed with a default ranking list size of 1000. This is the standard size in TREC experiments for the MAP calculation. As it can be seen the obtained correlation coefficients, among the different retrieval models, are similar. This similarity, in the obtained results with the

²Topic 672 has been removed since no relevant documents can be found for it in the collection.

³The retrieval models have been tested with the Terrier search engine developed by *The University of Glasgow*.

⁴Obtained results are statistically significant at a level of 0.01.

	BM25			LM			PL2		
	Sp	Pe	Ke	Sp	Pe	Ke	Sp	Pe	Ke
Min	-0.53	-0.49	-0.37	-0.54	-0.42	-0.38	-0.53	-0.45	-0.37
σ	0.49	0.39	0.34	0.48	0.35	0.33	0.43	0.30	0.29
σ_{max}	0.57	0.40	0.41	0.54	0.40	0.39	0.52	0.37	0.37

Table 1: Spearman(Sp), Pearson(Pe) and Kendall (Ke) correlation coefficients obtained with the proposed measures.

different retrieval models, can be interpreted as a proof of a common behaviour in terms of scoring, as was suggested before.

Minimum score measure shows a significant correlation degree, although its performance could drop when a whole ranking list is used. This is a consequence of the ranking list trend to achieve zero at the lower positions, when the scores have been normalised by the maximum score found. As was expected and due of the *ranking list tail* effect described before, standard deviation shows a poor performance measuring the scores dispersion.

On the other hand the results obtained with the maximum standard deviation outperforms to those achieved with standard deviation. Therefore σ_{max} avoids, at least in part, the lack of precision, in terms of dispersion measurement, obtained by the classic standard deviation.

5. CONCLUSIONS

In this paper some novel query performance predictors have been introduced. The obtained results show that measures based on standard deviation over scores ranking list, can be used to predict the quality of a search system reply.

The application of standard deviation as a dispersion measure for a ranking list, has shown to be a weak approach due to the noise introduced by the not relevant documents set retrieved. In order to avoid the prior effect the σ_{max} has been applied and it has improved the results acting as a noise reduction method. The correlation degree has been calculated with the most important correlation coefficients obtaining for all of them similar results. The obtained results outperform pre-retrieval approaches with a similar computational cost. In relation with post-retrieval approaches the results obtained are similar with the advantage of a much lower computational cost.

The consistency of the obtained results with the different retrieval models suggests the validity of the proposed method independently of the retrieval model. The main reason of this generalisation is achieved due it is based on a common expected behaviour of any ranking retrieval model, that is, the ability to distinguish between relevant and not-relevant documents for a specific topic.

6. FUTURE WORK

As a consequence of the study of the related work and the development of a new family of predictors, some open questions have been identified which could guide further research about this topic. Firstly it has not been established clearly which correlation coefficient is more adequate. Evaluations based on any of the three traditional correlation coefficients can be found, these are applied indistinctly without a clear justification of its use. This issue has been recently discussed by Claudia Hauff et al. in [7].

Moreover an evaluation based only on correlation coefficients can be hard to interpret. As can be found on the related literature, the obtained results in terms of correlation degree between prediction methods are almost equivalent for many of them. Some other measures have been proposed as *Kendall Average Precision* by Yilmaz et al. [15], and *Root Mean Square Error* in [7]. In our opinion a new family of different evaluation measures should be proposed, these measures should be more focused on the qualitative aspects of the prediction methods, and thus it should guide us towards a better understanding of the possible applications of prediction methods.

Finally, the application of prediction methods to improve some of the typical tasks related with information retrieval, as query expansion or search systems responses selection, should be further studied. We will try to answer the open questions raised on this work in further research.

Bibliography

- [1] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Query difficulty, robustness, and selective application of query expansion. In *ECIR*, pages 127–137, 2004.
- [2] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.
- [3] Javed A. Aslam and Virgil Pavlu. Query hardness estimation using jensen-shannon divergence among multiple scoring functions. *Advances in Information Retrieval*, 4425:198–209, 2007.
- [4] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 390, New York, 2006. ACM Press.
- [5] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, New York, 2002. ACM Press.
- [6] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Precision prediction based on ranked list coherence. *Information Retrieval.*, 9(6):723–755, 2006.
- [7] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. *The Combination and Evaluation of Query Performance Prediction Methods*. 2009.
- [8] Claudia Hauff, Vanessa Murdock, and Ricardo Baeza-Yates. Improved query difficulty prediction for the web. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, page 439, 2008.
- [9] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval*, pages 43–54, 2004.
- [10] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, January 2001.
- [11] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808, 2000.
- [12] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 267–275, 2001.
- [13] Stephen Robertson. On score distributions and relevance. *Advances in Information Retrieval*, 4425:40–51, 2007.
- [14] Ellen M. Voorhees. Overview of the trec 2004 robust retrieval track. In *In Proceedings of the Thirteenth Text REtrieval Conference (TREC)*, 2004.
- [15] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA, 2008. ACM.
- [16] Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. Learning to estimate query difficulty. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 512, 2005.
- [17] Ying Zhao, Falk Scholer, and Yohannes Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. *Advances in Information Retrieval*, 4956:52–64, 2008.
- [18] Yun Zhou and W. Bruce Croft. Query performance prediction in web search environments. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 543, 2007.