Ranking List Dispersion as a Query Performance Predictor*

Joaquín Pérez-Iglesias and Lourdes Araujo

Universidad Nacional de Educación a Distancia
Madrid 28040, Spain
joaquin.perez@lsi.uned.es, lurdes@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 hypothesis 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 use of measures based on standard deviation of ranking list scores, as a prediction value, shows a significant correlation degree in terms of average precision.

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. The proposed method for query performance prediction 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. This work is based on the hypothesis that different scores distributions for good and poor performance queries can be observed.

Related approaches that use ranking list scores can be found in the works carried out by Diaz [1], where the similarity between the scores of topically close documents, is applied as a prediction value. A similar approach was proposed by Vinay [2], in this case the prediction is based on the correlation between the actual rank and a computed expected rank, where the expected rank is obtained modelling the score of a document as a Gaussian random variable.

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. A search system ranks the related

* This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and the Spanish MICINN project TIN2007-68083.

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 371–374, 2009.

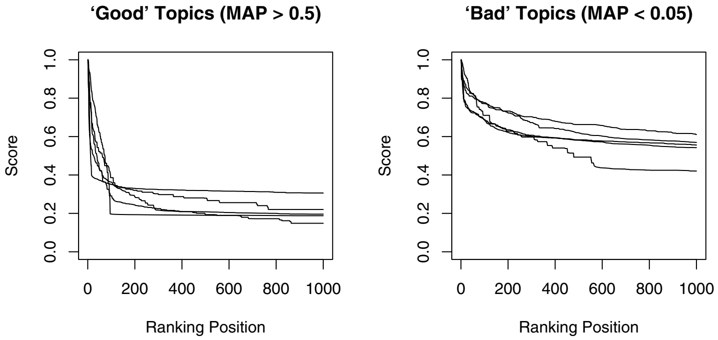

Fig. 1. 5 Best Performing Topics (left) Vs 5 Worst Performing Topics (right), from Robust 2004 using BM25. Scores have been normalised in $[0, 1]$. The maximum number of retrieved documents has been fixed to 1000.

documents found within the collection. For this purpose a ranking function assigns a weight (or score) to each document in the collection. In a 'naive' sense the scores can be interpreted as 'quantitative measures' of the documents 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.

Differences in terms of scores dispersion can be observed in figure 1 for the topics that achieve the best performance and those that obtain the lowest values in terms of AP (Average Precision) for Robust 2004 [3].

2.1 Proposed Measures

In this work we have tested different approaches to capture and measure dispersion along the obtained ranking list. Some prior studies have tried to model how document weights are distributed along a ranking list. In general, 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 [4,5]. Generally a majority 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. As a consequence, a ranking list shape holds a long *tail* where a majority of not relevant documents are placed.

Some notation is needed to define the next 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, placed at position i into the ranking list, is defined as $score(d_i)$.

Standard Deviation: Given ranking list scores mean $\mu(RL)$, standard deviation is computed as next:

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

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 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.

Maximum Standard Deviation: In order to minimise the effect of low scores high frequency, the maximum standard deviation is proposed. This estimator is based on the idea of computing the standard deviation at each point in the ranking list, and selecting the maximum value.

$$\sigma_{max} = max[\forall d \in RL, \sigma(RL_{[1,d]}))] \tag{2}$$

Standard Deviation at *k*: Standard deviation measured at a cut point k of the ranking list (σ_k). With the selection of a suitable k value, the noise introduced by low scores is removed. The k value is fixed at the ranking position that maximise the correlation degree with AP.

3 Results and Conclusions

The different measures proposed in this paper has been tested with the set of documents from TREC Disk4 & 5, minus Congressional Record and the topics used in the Robust 2004 track[1]. Only the field title from topics has been employed in the experiments. We have selected three well-known retrieval models (BM25,LM and PL2) to test the validity and compare the obtained prediction values among them.

The obtained results[2] appears in table 1. These experiments were executed with a default ranking list size of 1000, this was the default number of documents employed for the calculation of MAP in Robust 2004.

As can be seen the obtained correlation coefficients, with the same measure, for different retrieval models are similar. As it was expected a common behaviour for the proposed retrieval models can be observed.

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

[2] The correlation coefficients obtained are statistically significant at a level of 0.01.

In relation with the ability of the proposed measures to capture dispersion, the best results have been obtained with the selection of an optimal ranking list size k for σ_k. The size of the ranking list that maximises the correlation for all retrieval models is 100. Opposite to this, standard deviation exhibits a worse performance than the rest of measures as was affected by the described problem of the *ranking list tail.* 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.

Table 1. Pearson and Kendall correlation coefficients obtained with the proposed measures for different retrieval models. Strongest correlation values appear in bold.

	BM25		LM		PL2	
	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
σ	0.39	0.34	0.35	0.33	0.30	0.29
σ_{max}	0.40	0.41	0.40	0.39	0.37	0.37
σ_{100}	**0.55**	**0.41**	**0.53**	**0.41**	**0.53**	**0.39**

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. Further research in the selection of a suitable cut point for measuring the standard deviation, should be carried out to improve the obtained results.

References

1. Diaz, F.: Performance prediction using spatial autocorrelation. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 2007, p. 583. ACM Press, New York (2007)
2. Vinay, V., Milic-Frayling, N., Cox, I.: Estimating retrieval effectiveness using rank distributions. In: CIKM 2008: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 1425–1426. ACM, New York (2008)
3. Voorhees, E.M.: Overview of the trec 2004 robust retrieval track. In: Proceedings of the Thirteenth Text REtrieval Conference, TREC (2004)
4. Manmatha, R., Rath, T., Feng, F.: Modeling score distributions for combining the outputs of search engines. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 2001, pp. 267–275 (2001)
5. Robertson, S.: On score distributions and relevance. Advances in Information Retrieval 4425, 40–51 (2007)