# Improved Query Performance Prediction Using Standard Deviation

Ronan Cummins
School of Computing Science
University of Glasgow
Scotland
ronan.cummins@nuigalway.ie

Joemon M. Jose
School of Computing Science
University of Glasgow
Scotland
jj@dcs.gla.ac.uk

Colm O'Riordan
Department of IT
National University of Ireland, Galway
colmor@it.nuigalway.ie

## ABSTRACT

Query performance prediction (QPP) is an important task in information retrieval (IR). In this paper, we (1) develop a new predictor based on the standard deviation of scores in a variable length ranked list, and (2) we show that this new predictor outperforms state-of-the-art approaches without the need for tuning.

**Categories and Subject Descriptors:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval: Query formulation

**General Terms:** Experimentation, Measurement, Performance

**Keywords:** Information Retrieval, Query Performance Prediction

## 1. INTRODUCTION

Query performance prediction (QPP) has been a vibrant area of IR research over the last decade [2, 4, 3]. The motivation for QPP is that, if we can predict the performance of a query for a given system, we can automatically develop different strategies for dealing with these different queries. Predictors for this task are usually divided into two classes: *pre-retrieval* and *post-retrieval*. Pre-retrieval predictors are usually computationally less expensive but suffer from poor performance. Post-retrieval predictors are more computationally expensive as they use the ranked output (and/or scores) of a system, but achieve a higher performance than their counterparts. In general, the effectiveness a predictor is usually measured by calculating the correlation between the output of the predictor and the actual performance (i.e. average precision) of the queries on a system. Pearson's ($r$) and Spearman's ($\rho$) are two common correlation coefficient's used.

## 2. TEST COLLECTIONS

The data used in this paper consists of a number of TREC collections and a considerably large number of topics available for those collections. The *title* field was used as a short query for each of the collections, while the *desc* field was used as another set of queries[1].Table 1 shows details of the data that consists of over 500 different topics.


*SIGIR'11,* July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

**Table 1: News (top) and Web (bottom) Collections**

| Collection | Documents | Topic Range | # Topics | length | |
|---|---|---|---|---|---|
| | | | | title | desc |
| AP | 242,918 | 051-200 | 149 | 2.4 | 7.5 |
| FBIS | 130,471 | 301-450 | 116 | 2.4 | 7.6 |
| FT | 210,158 | 250-450 | 188 | 2.5 | 7.6 |
| WSJ | 130,837 | 051-200 | 150 | 2.4 | 7.5 |
| LA | 131896 | 301-450 | 143 | 2.4 | 7.6 |
| OHSU | 293,856 | 001-63 | 63 | - | 6.7 |
| WT2G | 221,066 | 401-450 | 50 | 2.4 | 6.5 |
| WT10G | 1,692,096 | 451-550 | 100 | 2.4 | 6.5 |

## 3. STANDARD DEVIATION FOR QPP

Recent research has shown that the standard deviation ($\sigma$) of scores in a ranked list is a good predictor of query performance [4]. The intuition is that, a good query is one for which the scores of documents at the head of the ranked-list are highly dispersed (i.e. the user has chosen good query terms that enhance the signal of a certain number of topical documents compared to the noise of the collection). Some standard approaches [4] have shown that the standard deviation at fixed cut-off points (e.g. 100 documents) is correlated with query performance. It has also been shown [4] that even better prediction can be obtained if a variable cut-off point is used (i.e. a different cut-off point for each query) using a tuning parameter. We adopt this idea and derive a simple, yet intuitive, method of automatically determining the cut-off value for each query.

**Table 2: Correlation of $\sigma_{x\%}$ with average precision**

| AP (title) | $\sigma_{90\%}$ | $\sigma_{75\%}$ | $\sigma_{60\%}$ | $\sigma_{50\%}$ | $\sigma_{40\%}$ | $\sigma_{25\%}$ | $n(\sigma_{50\%})$ |
|---|---|---|---|---|---|---|---|
| Pearson ($r$) | 0.352 | 0.421 | 0.535 | **0.624** | 0.617 | 0.505 | 0.672 |
| Spearman ($\rho$) | 0.312 | 0.348 | 0.500 | 0.602 | **0.617** | 0.542 | 0.650 |
| OHSU (desc) | $\sigma_{90\%}$ | $\sigma_{75\%}$ | $\sigma_{60\%}$ | $\sigma_{50\%}$ | $\sigma_{40\%}$ | $\sigma_{25\%}$ | $n(\sigma_{50\%})$ |
| Pearson ($r$) | 0.232 | 0.328 | 0.481 | **0.570** | 0.516 | 0.299 | 0.622 |
| Spearman ($\rho$) | 0.323 | 0.335 | 0.475 | **0.535** | 0.534 | 0.347 | 0.538 |
| WT2G (title) | $\sigma_{90\%}$ | $\sigma_{75\%}$ | $\sigma_{60\%}$ | $\sigma_{50\%}$ | $\sigma_{40\%}$ | $\sigma_{25\%}$ | $n(\sigma_{50\%})$ |
| Pearson ($r$) | 0.071 | 0.343 | 0.433 | 0.536 | **0.621** | 0.359 | 0.590 |
| Spearman ($\rho$) | 0.045 | 0.373 | 0.380 | **0.526** | 0.525 | 0.331 | 0.556 |

As it is the head of the retrieval list that is important, we calculate the standard deviation of the scores of the first $N$ documents, where $N$ is the number of documents that

[1]For the OHSUMED collection only the *desc* was used, as it is the actual information need for the topic

**Table 3: Natural tendency for longer queries to return increased $\sigma$ of scores without an increase in performance (MAP)**

| | title | | desc | |
|---|---|---|---|---|
| | $MAP$ | $avg(\sigma_{50\%})$ | $MAP$ | $avg(\sigma_{50\%})$ |
| AP | 0.159 | 1.811 | 0.151 | 2.597 |
| FBIS | 0.225 | 1.839 | 0.202 | 2.567 |
| FT | 0.228 | 1.983 | 0.219 | 2.739 |
| WSJ | 0.221 | 1.924 | 0.209 | 2.796 |
| WT2G | 0.224 | 1.847 | 0.227 | 2.626 |

are assigned a score greater than a certain percentage ($x$) of the top score. For example, if we choose $x = 90\%$, all documents that have a score of at least 90% of the top score are included in the standard deviation calculation. Table 2 shows the performance of this approach on three of the collections for a $BM25$ system. We can see that performance (i.e. correlation) is optimised at $x = 50\%$ (i.e. all document scores that are at least 50% of the top score for a given query are are included in the standard deviation calculation). Results on all other collections used in this work (not included due to space limitations) report a similar trend. This simple method means that a varying number of documents are included in the standard deviation calculation, and that these documents are of a certain quality (as determined by the system itself).

Furthermore, we also determined that there is a natural tendency for longer queries to produce ranked lists with a higher deviation of document score, although these longer queries might not produce a higher performance. Table 3 outlines this phenomenon. Therefore, we normalised the standard deviation with respect to query length. Thus, our new normalised query performance predictor is $n(\sigma_{50\%}) = \frac{\sigma_{50\%}}{sqrt(ql)}$ where $ql$ is the query length. The last column of Table 2 confirms that this new normalised predictor outperforms the unnormalised version on the collections. Furthermore, both new predictors ($\sigma_{50\%}$ and $n(\sigma_{50\%})$) are significantly correlated with average precision. Now that we have developed a new predictor we compare it against some state-of-the-art approaches.

# 4. EXPERIMENTS

In these experiments, we use a $BM25$ system and compare the performance of a number of state-of-the-art predictors against our newly developed predictor. The best pre-retrieval predictors from the literature are the simplified clarity score ($scs$), the average $idf$ of query terms ($idf_{avg}$), and the maximum $idf$ of the query terms ($idf_{max}$). The best post-retrieval predictors from the literature are query clarity ($clarity$) [1], $ncq$ [5], standard deviation at 100 documents ($\sigma_{100}$), the maximum standard deviation in the ranked-list ($\sigma_{max}$), and a variable cut-off point ($k$) approach [4] ($\sigma_k$) which includes a tuning parameter $\lambda$ which we set to 5.

## 4.1 Performance Comparison

Table 4 shows the performance of the predictors averaged over the News collections for each query type (*title* and *desc*). Firstly, we can see that while pre-retrieval predictors are useful for short queries, they are poor on longer queries. The clarity score achieves steady performance across the collections and query types. However, the predictors based on standard deviation are generally more highly correlated with query performance. Table 5 shows the best predictors on larger Web collections. There is a significant correlation with average precision on all the individual collections for the post-retrieval predictors which is mainly due to the large number of queries we use for each collection. The new predictor $n(\sigma_{50\%})$ outperforms the other predictors consistently over all query types and collections. Simply to outline the consistency of the increases over a good baseline, we performed a paired Wilcoxon test on the 15 (7 *title* sets and 8 *desc* sets) $\rho$ coefficients of $n(\sigma_{50\%})$ compared to $ncq$ and determined that the p-value was 0.012.

**Table 4: Correlation coefficients ($r$ and $\rho$) averaged for the News collections for title and desc queries**

| | title | | desc | |
|---|---|---|---|---|
| Predictor | $avg(r)$ | $avg(\rho)$ | $avg(r)$ | $avg(\rho)$ |
| $scs$ | 0.374 | 0.307 | 0.205 | 0.172 |
| $idf_{max}$ | 0.332 | 0.295 | 0.191 | 0.208 |
| $idf_{avg}$ | 0.423 | 0.344 | 0.250 | 0.221 |
| $clarity$ | 0.381 | 0.417 | 0.345 | 0.379 |
| $\sigma_{100}$ | 0.456 | 0.442 | 0.499 | 0.504 |
| $\sigma_{max}$ | 0.475 | 0.493 | 0.404 | 0.406 |
| $\sigma_{k^5}$ | 0.448 | 0.338 | 0.281 | 0.254 |
| $ncq$ | 0.523 | 0.429 | 0.527 | 0.506 |
| $\sigma_{50\%}$ | 0.501 | 0.487 | 0.535 | 0.525 |
| $n(\sigma_{50\%})$ | **0.569** | **0.538** | **0.604** | **0.588** |

**Table 5: Spearman correlation ($\rho$) for best predictors on Web collections**

| Collection | $clarity$ | $\sigma_{100}$ | $ncq$ | $\sigma_{50\%}$ | $n(\sigma_{50\%})$ |
|---|---|---|---|---|---|
| WT2G (title) | 0.352 | 0.445 | 0.411 | 0.502 | **0.531** |
| WT2G (desc) | 0.321 | 0.585 | 0.593 | 0.567 | **0.606** |
| WT10G (title) | 0.358 | 0.356 | 0.342 | **0.447** | 0.423 |
| WT10G (desc) | 0.401 | 0.502 | 0.492 | 0.550 | **0.566** |

# 5. CONCLUSION

In this paper, we have developed a new post-retrieval predictor for query performance, that needs no tuning to achieve a high correlation with average precision. The new predictor outperforms state-of-the-art predictors on a number of test collections for both short and medium length queries. The predictor is intuitively simple and less computationally expensive than some other approaches, such as the *clarity* score.

# Acknowledgments

The first author is funded by the Irish Research Council for Science, Engineering and Technology (IRCSET), co-funded by Marie Curie Actions under FP7.

# 6. REFERENCES

[1] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR '02*, pages 299–306, New York, NY, USA, 2002. ACM.

[2] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A survey of pre-retrieval query performance predictors. In *CIKM*, pages 1419–1420, 2008.

[3] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE*, pages 43–54, 2004.

[4] Joaquín Pérez-Iglesias and Lourdes Araujo. Standard deviation as a query hardness estimator. In *SPIRE*, pages 207–212, 2010.

[5] Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *ICTIR*, pages 305–312, 2009.