Query Performance Prediction

Your name

ABSTRACT

Write your abstract here. Your abstract should *concisely* say (i) *why* the topic is interesting, (ii) *what* you do in your study, (iii) *how* you did your study and (iv) *what* the results were

1. INTRODUCTION

Write your introduction here. Get inspiration on how to structure and formulate an introduction from the studies you review. Make sure you describe what query performance prediction is and why it is useful.

Several studies on query-performance prediction have been published before and after [1]. In your introduction, give an overview of ***at least three*** studies of query-performance prediction published by the ACM between 2005 and 2016. You may find these studies by searching e.g. the ACM digital library (http://dl.acm.org) or Google Scholar. Your literature review of the 3+ papers must (i) describe the method proposed in the paper, how the method was evaluated and what the results were. Furthermore, it should be clear how each paper differ from the other paper you review. **Remember: the literature review is meant to help the reader understand where there is a gap in the existing research that you can fill**. Therefore, select papers that are as close as possible to [1].

You *must* cite your sources when/if you use a specific phrasing. Failure to do so will be considered plagiarism. aaaaa

2. DATA SETS & QUERIES

List in a table similar to Table 1 in [1], for each data set, the following characteristics:

1. Name
2. Number of documents
3. Average document length
4. Minimum document length
5. Maximum document length

Furthermore, list the number of queries and the average query length for the superset of the queries.

3. EXPERIMENTS

For indexing, retrieval and evaluation use INDRI and use TREC_EVAL. Index the data sets using the Krovetz stemmer and stop word removal using the list http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words. Use the BM25 ranking model for retrieval. You are not required to tune any parameter, but are welcome to do so.

Use values of $N = \{20, 50, 100, 200\}$ and report a table similar to that of Table 2 in [1] setting $\sigma = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and include the normalised version (see the paper). Use both Pearson's r and Spearman's ρ, and use only the title of each query. Where is the correlation with P@10 the highest for each N? Where is it best overall? Why do you think that is? Repeat the above with at least one other metric (e.g. MRR, nDCG etc.). Which one gives you the highest correlation?

As the standard deviation assumes data are normally distributed (which is not case for many real-life data sets), repeat the above analysis, but instead of σ use the *mean absolute deviation*:

$$\text{MAD} = \text{median}\left(x_i - \text{median}(x)\right) \tag{1}$$

Does MAD correlate with P@10 better than compared to using the standard deviation? Why? What is the best correlation you get? What about the normalised version of the MAD?

Finally, produce a Table similar to Table 4 in [1] for one of the datasets of your choosing. Use the simplified clarify score (scs), average IDF (idf_{avg}), query scope (qs) and $\sigma_{\max}$ predictors.

The simplified clarify score [**?**] (scs) is given by:

$$scs(Q) = \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{\text{coll}(w)}} \tag{2}$$

where w is a word in the vocabulary V of the data set or query Q, $P(w|Q)$ is the conditional probability (relative frequency or MLE estimate) of w in the query Q and $P_{\text{coll}}(w)$ is the probability of w in the data set (the relative frequency of w in V).

The query scope [**?**] (qs) is given by:

$$qs(Q) = -\log(n_Q/N) \tag{3}$$

where log denotes the natural logarithm, n_q is the number of documents in the data set that contains *at least* one of the query terms, and N is the number of documents in the data set.

Compare your best correlations with these predictors.

4. CONCLUSION

Write your conclusion here

5. REFERENCES

[1] R. Cummins, J. Jose, and C. O'Riordan. Improved query performance prediction using standard deviation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1089–1090. ACM, 2011.