

Query Quality: User Ratings and System Predictions

Claudia Hauff¹, Franciska de Jong¹, Diane Kelly², Leif Azzopardi³
{c.hauff, f.m.g.dejong}@ewi.utwente.nl, dianek@email.unc.edu, leif@dcs.gla.ac.uk

UNIVERSITY OF TWENTE.



¹University of Twente, The Netherlands
²University of North Carolina, Chapel Hill, USA
³University of Glasgow, United Kingdom



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

1. Introduction

- **Research Question:** Do user ratings of query quality align with predictions of system effectiveness made by automatic query performance prediction (QPP) methods?
- **Motivation:**
 - User perceptions of the quality of query suggestions affect their view of a search system's effectiveness.
 - QPP methods rely on “rules of thumb” of how users judge a query's quality. Do QPP actually reflect the intuitions of human assessors?

2. Previous Work

- **Explicit user ratings versus performance [2]:**
 - IR experts rated queries for a well-known newswire corpus as easy/medium/hard without viewing the search results.
 - Low level of correct ratings and inter-rater agreement.
- **Inferred user ratings versus performance [1, 3]:**
 - Time to find a relevant document as implicit rating.
 - Weak correlations between implicit ratings and pre/post-retrieval QPP methods.
- **In this poster:**
 - A more realistic data set.
 - Explicit user ratings of query quality.

3. Empirical Study

- **Data Set:**
 - ClueWeb09 (cat. B) corpus with ≈ 50 million documents.
 - Preprocessing: stopword removal & Porter stemming.
 - Retrieval approach: Language Modeling with Dirichlet smoothing.
 - Fifty topics (consisting of *query* and *description*) from the TREC 2009 Web adhoc retrieval task.
 - E.g. topic 3 consists of the *query* “getting organized” and the *description* “Find tips, resources, supplies for getting organized and reducing clutter”.
- **Human Assessors of Query Quality:**
 - Eighteen post-graduate computer science students (expert users).
 - For each topic, the assessors were asked to judge the *query's* quality on a scale from 1 (poor quality) to 5 (high quality), given the *description*, but **without** viewing the search results.
- **Query Performance Prediction Methods:**
 - Pre-retrieval (parameter-free):
 - Maximum Inverse Document Frequency (*MaxIDF*).
 - Summed Term Weight Variability (*SumVAR*).
 - Summed Collection Query Similarity (*SumSCQ*).
 - Post-retrieval (used in their best parameter settings):
 - Clarity Score.
 - Query Feedback.
 - Query Commitment.

4. Assessor Ratings

- Topic set partitioned according to query effectiveness (Avg. Precision, Precision at 30 documents).
- Assessors on average able to recognize well performing queries (avg. inter-rater agreement: $\kappa = 0.36$).

Query Partitions	Performance		Assessor Ratings	
	AP	P@30	AP	P@30
1-10 (best)	0.414	0.629	3.87 (1.07)	4.00 (1.01)
11-20	0.298	0.470	3.72 (1.09)	3.53 (1.20)
21-30	0.099	0.272	3.24 (1.37)	3.31 (1.29)
31-40	0.032	0.133	2.79 (1.20)	2.89 (1.33)
41-50 (worse)	0.005	0.038	2.51 (1.48)	2.40 (1.34)

5. QPP Methods vs. System Effectiveness & Assessor Ratings

- Table: Kendall's Tau rank correlation between QPP methods and (i) system effectiveness, as well as (ii) assessor ratings.
- Pre-retrieval QPP methods are better predictors of system effectiveness, possibly due to the noisy nature of ClueWeb09.
- User ratings are most highly correlated with pre-retrieval QPP methods; the correlations are moderate at best.

Pre/Post Ret. Predictors	Performance		Assessor Ratings		
	AP	P@30	Min	Avg	Max
<i>MaxIDF</i>	0.35†	0.19	-0.09	0.09	0.29†
<i>SumSCQ</i>	0.39†	0.35†	0.20	0.31†	0.49†
<i>SumVAR</i>	0.42†	0.38†	0.17	0.28†	0.43†
<i>Clarity Score</i>	0.27†	0.18	-0.10	0.02	0.19
<i>Query Feedback</i>	0.37†	0.29†	0.12	0.28†	0.44†
<i>Query Commit.</i>	0.26†	0.11	-0.15	0.01	0.18

† statistically significant correlation ($p < 0.01$)

6. Conclusions

- The relationship between system predictions (QPP) and explicit user ratings of query quality were investigated.
- Assessor ratings are more often significantly correlated with pre-retrieval than with post-retrieval QPP methods.
- Overall, the found correlations were weak to moderate, making QPP methods poor proxies for user ratings.

References

- [1] A. Turpin and W. Hersh. Do clarity scores for queries correlate with user performance? In *ADC '04*, pp. 85–91, 2004
- [2] E. Voorhees and D. Harman. Overview of the sixth Text REtrieval Conference. In *Proceedings of TREC-6*, 1997.
- [3] Y. Zhao and F. Scholer. Predicting query performance for user-based search tasks. In *ADC '07*, pp. 112–115, 2007.