An Enhanced Evaluation Framework for Query Performance Prediction

Guglielmo Faggioli[1], **Oleg Zendel[2]**, J Shane Culpepper[2], Nicola Ferro[1] and Falk Scholer

[1]University of Padova, Padova, Italy

[2]RMIT University, Melbourne, Australia

March 29, 2021

ECIR 2021 LUCCA | TUSCANY ONLINE EVENT

Outline

- QPP – definition and motivation
- Existing evaluation in QPP
- Research questions
- Related work
- Limitations in current evaluation
- Proposed solution
- Experiments

Query Performance Prediction

Estimating the effectiveness of a search performed in response to a query without relevance information.

Pre-retrieval predictors:

Analyze the query and corpus statistics prior to retrieval.

Post-retrieval predictors:

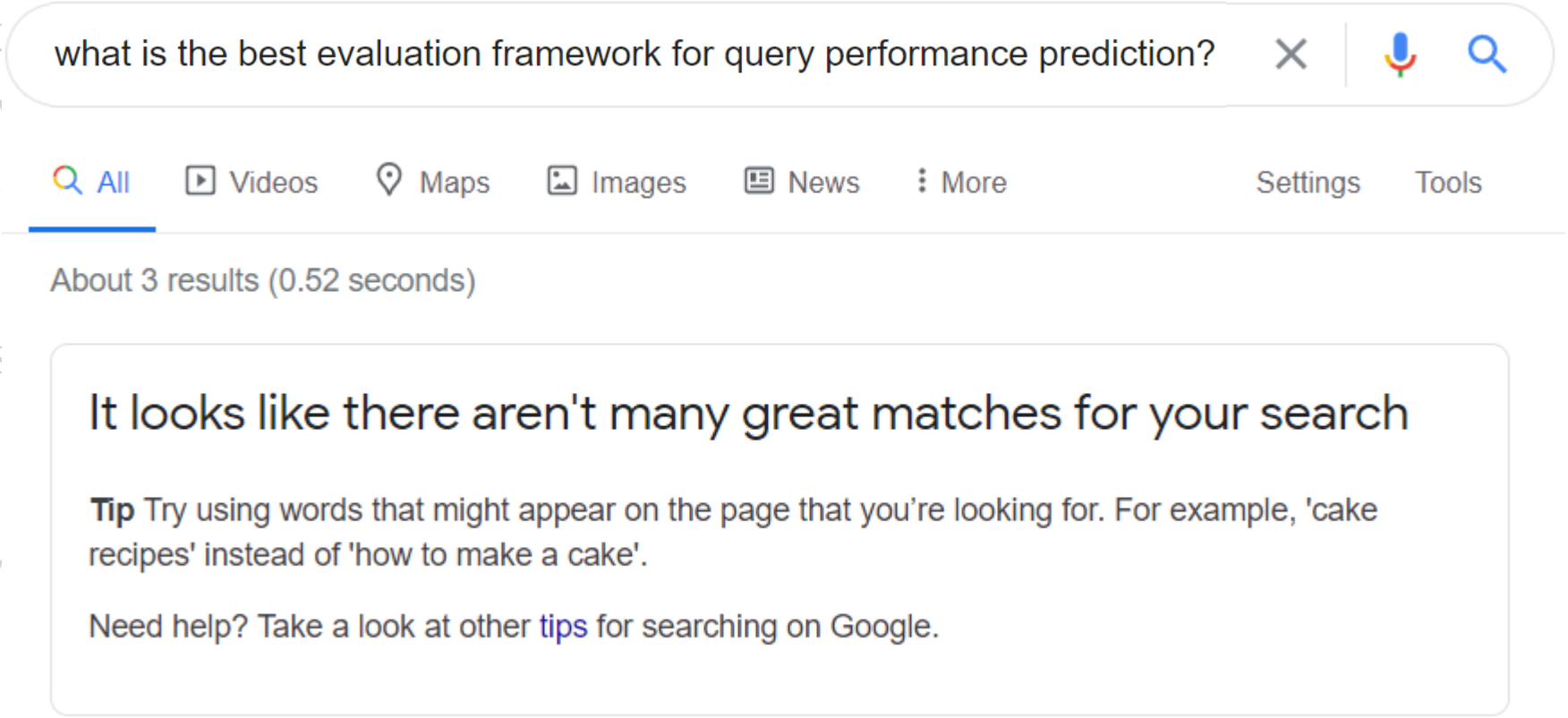
Analyze information induced from top retrieved documents.

Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus

Motivation – Potential Applications

- Feedback to the user
- Feedbac
  - Selecti
  - Federa
  - Fusion
- Convers
- Query s
- Identify



Motivation – Potential Applications

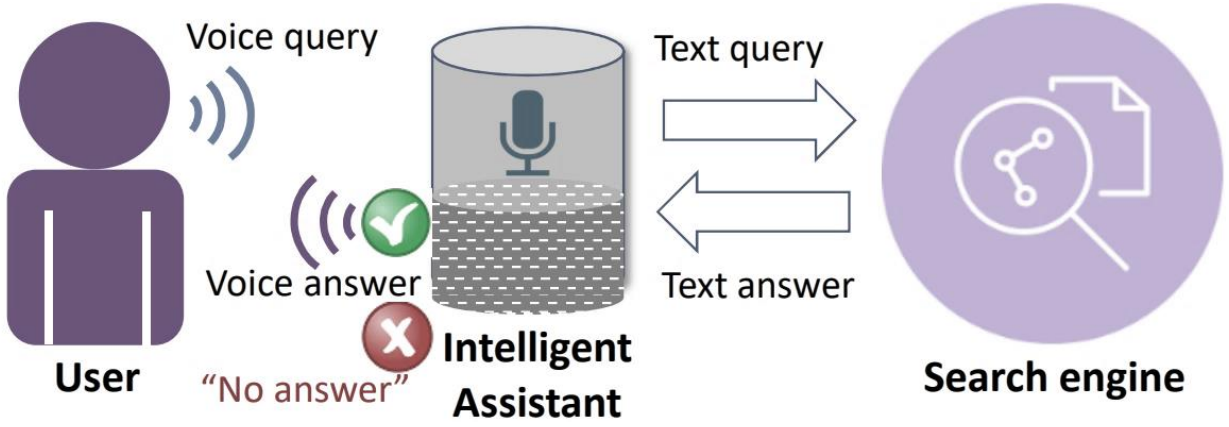
- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus

Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus

Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus



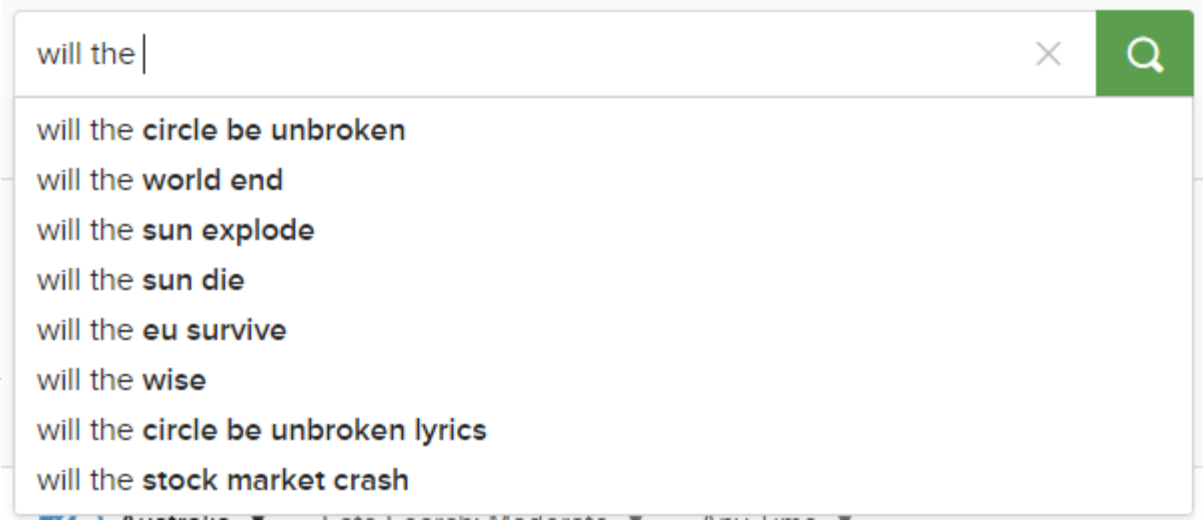
(Image credit: Roitman, H. et al., '19
"A Study of Query Performance Prediction for Answer Quality Determination")

Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus

Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing conter



Motivation – Potential Applications

- Feedback to the user
- Feedback to the system
  - Selective Query Expansion (SQE)
  - Federated search
  - Fusion
- Conversational search
- Query suggestions
- Identifying missing content in the corpus

Evaluation of Query Performance Prediction

Goal specific task	General task	Suitable evaluation measures
Identifying search failures / hard queries	Classification	Accuracy, AUC, AUCPR, F1-score
Predicting a retrieval effectiveness measure	Regression	R^2, MAE, MSE, RMSE
Ranking queries based on effectiveness / difficulty	Ranking	Pearson's r, Spearman's ρ, Kendall's τ, Kendall's dist., Spearman's footrule dist.

Existing Evaluation of QPP

- Existing evaluation relies on correlation measures.
- The most common are:
  - Pearson's r – linear correlation coefficient (parametric);
  - Spearman's ρ – monotonic rank correlation coefficient (non-parametric); and
  - Kendall's τ – monotonic rank correlation coefficient (non-parametric).
- Straightforward statistical significance testing for a single QPP method.
- The correlation is measured between the predicted values and the effectiveness measure.
- The most used effectiveness measure is AP.

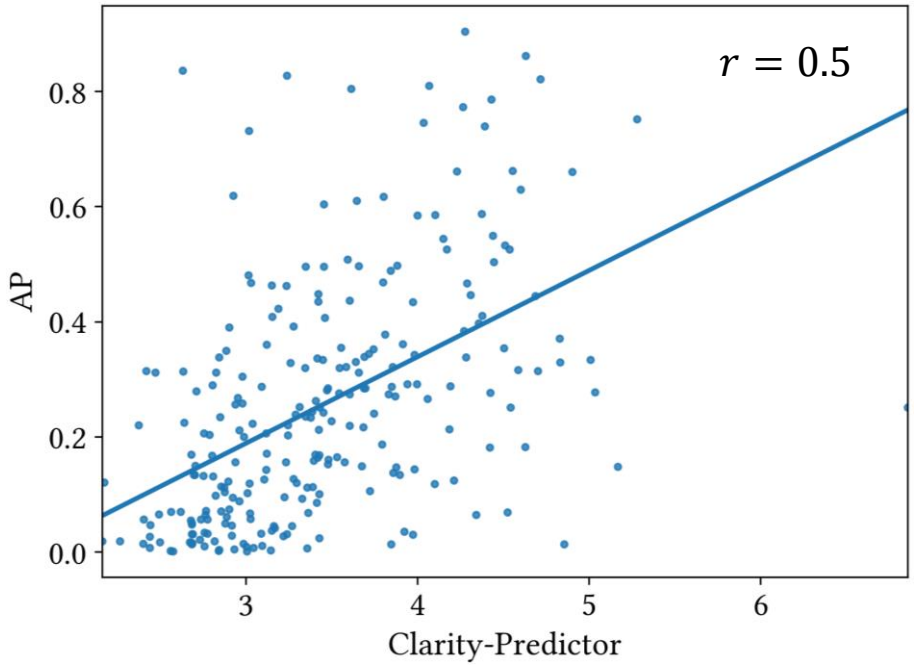
Existing Evaluation of QPP

- Existing evaluation relies on correlation measures.
- The most common are:
  - Pearson's r – linear correlation coefficient (parametric);
  - Spearman's ρ – monotonic rank correlation coefficient (non-parametric); and
  - Kendall's τ – monotonic rank correlation coefficient (non-parametric).
- Straightforward statistical significance testing for a single QPP method.
- The correlation is measured between the predicted values and the effectiveness measure.
- The most used effectiveness measure is AP.

Existing Evaluation of QPP

- Existing evaluation relies on correlation measures.
- The most common are:
  - Pearson's r – linear correlation coefficient (parametric);
  - Spearman's ρ – monotonic rank correlation coefficient (non-parametric); and
  - Kendall's τ – monotonic rank correlation coefficient (non-parametric).
- Straightforward statistical significance testing for a single QPP method.
- The correlation is measured between the predicted values and the effectiveness measure.
- The most used effectiveness measure is AP.

Existing Evaluation of QPP



- The correlation is measured between the predicted values and the effectiveness measure.
- The most used effectiveness measure is AP.

(Image: scatter plot with regression line, ROBUST Title queries)

Existing Evaluation of QPP (cont.)

- The correlation based evaluation method first mentioned in 1998.
  (E. M. Voorhees and D. K. Harman, '98)
- The correlation was adopted as the main measure in practice.
- It is suitable for a single predictor.

Existing Evaluation of QPP (cont.)

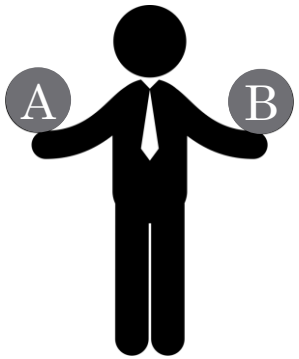
- The correlation based evaluation method first mentioned in 1998.
  (E. M. Voorhees and D. K. Harman, '98)
- The correlation was adopted as the main measure in practice.
- It is suitable for a single predictor.

Existing Evaluation of QPP (cont.)

- The correlation based evaluation method first mentioned in 1998.
  (E. M. Voorhees and D. K. Harman, '98)
- The correlation was adopted as the main measure in practice.
- It is suitable for a single predictor.

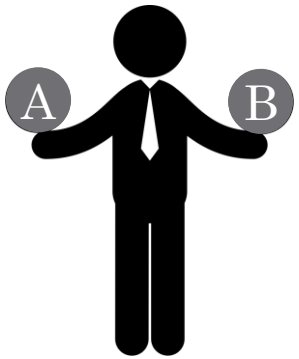
Existing Evaluation Comparing Multiple Methods

- The QPP method with higher correlation considered superior.
- For each predictor a distribution of values is achieved by resampling.
- Student's paired t-test is used for statistical testing.



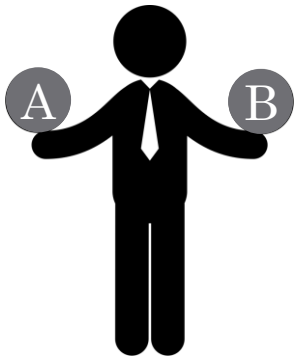
Existing Evaluation Comparing Multiple Methods

- The QPP method with higher correlation considered superior.
- For each predictor a distribution of values is achieved by resampling.
- Student's paired t-test is used for statistical testing.



Existing Evaluation Comparing Multiple Methods

- The QPP method with higher correlation considered superior.
- For each predictor a distribution of values is achieved by resampling.
- Student's paired t-test is used for statistical testing.



Research Questions

- What limitations exist in the current evaluation practices of query performance prediction?
- How can detailed statistical analysis and testing be applied to QPP evaluation exercises?
- What factors contribute to improving or reducing the performance of a QPP model?

Research Questions

- What limitations exist in the current evaluation practices of query performance prediction?
- How can detailed statistical analysis and testing be applied to QPP evaluation exercises?
- What factors contribute to improving or reducing the performance of a QPP model?

Research Questions

- What limitations exist in the current evaluation practices of query performance prediction?
- How can detailed statistical analysis and testing be applied to QPP evaluation exercises?
- What factors contribute to improving or reducing the performance of a QPP model?

Related Work

- The comparison of linear correlation coefficients should be based on Fisher's Z transformation (Meng, X.L. et al. , '92)
- The current correlation based evaluation can't be generalized for a predictor (Scholer, F. and Garcia, S., '09)
- Higher correlation does not necessarily attest to better prediction (Hauff, C. et al. '09)

Related Work

- The comparison of linear correlation coefficients should be based on Fisher's Z transformation (Meng, X.L. et al. , '92)
- The current correlation based evaluation can't be generalized for a predictor (Scholer, F. and Garcia, S., '09)
- Higher correlation does not necessarily attest to better prediction (Hauff, C. et al. '09)

Related Work

- The comparison of linear correlation coefficients should be based on Fisher's Z transformation (Meng, X.L. et al. , '92)
- The current correlation based evaluation can't be generalized for a predictor (Scholer, F. and Garcia, S., '09)
- Higher correlation does not necessarily attest to better prediction (Hauff, C. et al. '09)

Related Work (Cont.)

- Testing the difference of QPP methods with Fisher's Z transformation yields non significant results for many of the existing methods
  (Hauff, C. et al. '09)

- Relative QPPs prediction quality varies with respect to the effectiveness of queries used to represent the topics
  (Thomas, P. et al., '17, Zendel, O. et al., '19)

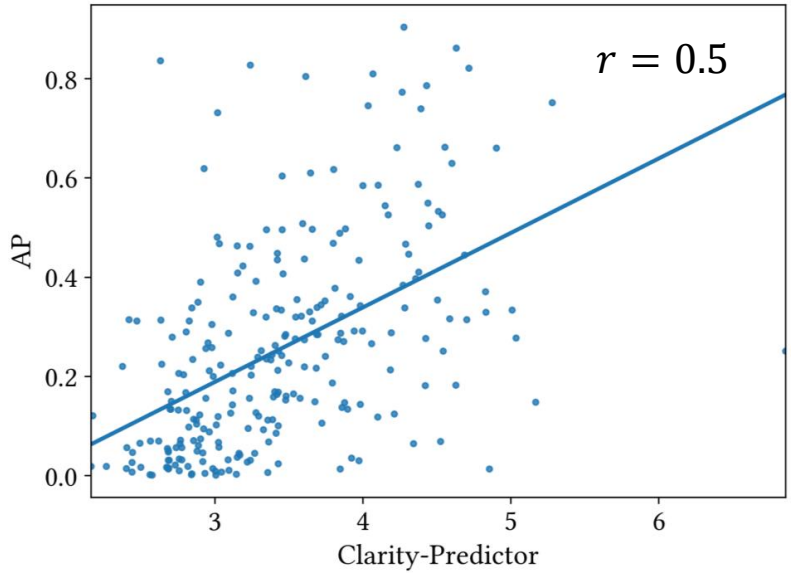
Related Work (Cont.)

- Testing the difference of QPP methods with Fisher's Z transformation yields non significant results for many of the existing methods
(Hauff, C. et al. '09)

- Relative QPPs prediction quality varies with respect to the effectiveness of queries used to represent the topics
(Thomas, P. et al., '17, Zendel, O. et al., '19)

Limitations in Current Evaluation

- Single aggregated value – hard to interpret
- Unable to identify hard queries – where did it fail
- Limited ability to analyse affect of different components
- Hard to generalize – very specific
- Unsound statistical testing



Limitations in Current Evaluation

- Single aggregated value – hard to interpret
- Unable to identify hard queries – where did it fail
- Limited ability to analyse affect of different components
- Hard to generalize – very specific
- Unsound statistical testing

Limitations in Current Evaluation

- Single aggregated value – hard to interpret
- Unable to identify hard queries – where did it fail
- Limited ability to analyse affect of different components
- Hard to generalize – very specific
- Unsound statistical testing

Limitations in Current Evaluation

- Single aggregated value – hard to interpret
- Unable to identify hard queries – where did it fail
- Limited ability to analyse affect of different components
- Hard to generalize – very specific
- Unsound statistical testing

Limitations in Current Evaluation

- Single aggregated value – hard to interpret
- Unable to identify hard queries – where did it fail
- Limited ability to analyse affect of different components
- Hard to generalize – very specific
- Unsound statistical testing

Proposed Solution

- Use a non-parametric association measure
- Model the prediction errors as a distribution over the queries
- Use well grounded statistical analyses

Proposed Solution

- Use a non-parametric association measure
- Model the prediction errors as a distribution over the queries
- Use well grounded statistical analyses

Proposed Solution

- Use a non-parametric association measure
- Model the prediction errors as a distribution over the queries
- Use well grounded statistical analyses

Scaled Absolute Rank Error (sARE_{AP})

$$\mathrm{sARE}_{AP}(q_i) := \frac{\left|r_i^p - r_i^e\right|}{|Q|}$$

q_i - query i

r_i^p - rank assigned by the predictor

r_i^e - rank assigned by the effectiveness measure

Q - set of queries

$$\mathrm{sARE}_{AP}(q_i) \in \left[0, \frac{n-1}{n}\right] \subseteq [0,1)$$

Scaled Mean Absolute Rank Error (sMARE_{AP})

$$\mathrm{sMARE}_{AP}(\mathcal{P}) \coloneqq \frac{1}{|Q|} \sum_{q_i \in Q} \mathrm{sARE}_{AP}(q_i)$$

q_i - query i

Q - set of queries

$$\mathrm{sMARE}_{AP}(\mathcal{P}) \in [0, 0.5]$$

Analysis Of Variance (ANOVA)

- Used to asses statistically significant differences among means.
- Model the observations in the form:

$$Data = Model + Error$$

$Data$ – explained variable (dependent variable)

$Model$ – the experimental factors (contains a coefficient for each factor)

$Error$ – the variance in the $Data$ that not explained by the $Model$

Analysis Of Variance (ANOVA)

- Used to asses statistically significant differences among means
- Model the observations in the form:

$$Data = Model + Error$$

$Data$ – explained variable (dependent variable)

$Model$ – the experimental factors (contains a coefficient for each factor)

$Error$ – the variance in the $Data$ that not explained by the $Model$

ANOVA Models

- Model without interactions between the factors (MD0)

$$y_{iqrs} = \mu + \tau_i + \gamma_q + \delta_r + \xi_s + \epsilon_{iqrs}$$

- Model with interactions between the factors (MD1)

$$y_{iqrs} = \mu + \tau_i + \nu_{j(i)} + \gamma_q + \cdots + (\tau\gamma)_{iq} + (\tau\delta)_{ir} + \cdots + \epsilon_{ijqrs}$$

ANOVA Models

- Model without interactions between the factors (MD0)

$$y_{iqrs} = \mu + \tau_i + \gamma_q + \delta_r + \xi_s + \epsilon_{iqrs}$$

- Model with interactions between the factors (MD1)

$$y_{iqrs} = \mu + \tau_i + \nu_{j(i)} + \gamma_q + \cdots + (\tau\gamma)_{iq} + (\tau\delta)_{ir} + \cdots + \epsilon_{ijqrs}$$

Post-Hoc Analysis

- Applied after the ANOVA to find pairwise statistical differences
- Adjusts for multiple comparisons (family-wise error rate)
- We apply Tukey's Honestly Significant Difference (HSD) test
- Tukey's HSD test is based on the studentized range distribution
- Defines confidence intervals for the differences

Post-Hoc Analysis

- Applied after the ANOVA to find pairwise statistical differences
- Adjusts for multiple comparisons (family-wise error rate)
- We apply Tukey's Honestly Significant Difference (HSD) test
- Tukey's HSD test is based on the studentized range distribution
- Defines confidence intervals for the differences

Post-Hoc Analysis

- Applied after the ANOVA to find pairwise statistical differences
- Adjusts for multiple comparisons (family-wise error rate)
- We apply Tukey's Honestly Significant Difference (HSD) test
- Tukey's HSD test is based on the studentized range distribution
- Defines confidence intervals for the differences

Post-Hoc Analysis

- Applied after the ANOVA to find pairwise statistical differences
- Adjusts for multiple comparisons (family-wise error rate)
- We apply Tukey's Honestly Significant Difference (HSD) test
- Tukey's HSD test is based on the studentized range distribution
- Defines confidence intervals for the differences

Post-Hoc Analysis

- Applied after the ANOVA to find pairwise statistical differences
- Adjusts for multiple comparisons (family-wise error rate)
- We apply Tukey's Honestly Significant Difference (HSD) test
- Tukey's HSD test is based on the studentized range distribution
- Defines confidence intervals for the differences

Experimental Setup

- Documents corpus:
  TREC ROBUST-04, ~528K documents, 249 TREC title queries
- Query variants:
  TREC Core 2017 - ROBUST-2004 corpus
  (Benham, R. and Culpepper, J.S., '17)
- Retrieval method:
  Query likelihood
  (Ponte, J.M. and Croft, W.B., '98)
- IR effectiveness measure:
  Average Precision (AP)

Experimental Setup (Cont.)

Post-retrieval	Pre-retrieval
Clarity (Cronen-Townsend, S. et al., '02)	SCQ, AvgSCQ, MaxSCQ (Zhao, Y. et al., '08)
NQC (Shtok, A. et al., '12)	Var, AvgVar, MaxVar (Zhao, Y. et al., '08)
WIG (Zhou, Y. and Croft, W.B., '07)	AvgIDF (Cronen-Townsend, S. et al., '04.)
SMV (Tao, Y. and Wu, S., '14)	MaxIDF (Scholer, F. et al., '04)
UEF (Shtok, A. et al., '10)	

- In total 16 different QPP methods

Experimental Setup (Cont.)

- 3 different stemmer configurations :
  Lovins, *Porter* and no stemming
- 5 different stoplist configurations:
  attire, *zettair*, *indri*, *lingpipe* and no stoplist
- Total of 15 retrieval pipelines
- Total of 240 QPP-system combinations

Experimental Setup (Cont.)

- 3 different stemmer configurations :
  Lovins, *Porter* and no stemming

- 5 different stoplist configurations:
  attire, *zettair*, *indri*, *lingpipe* and no stoplist

- Total of 15 retrieval pipelines
- Total of 240 QPP-system combinations

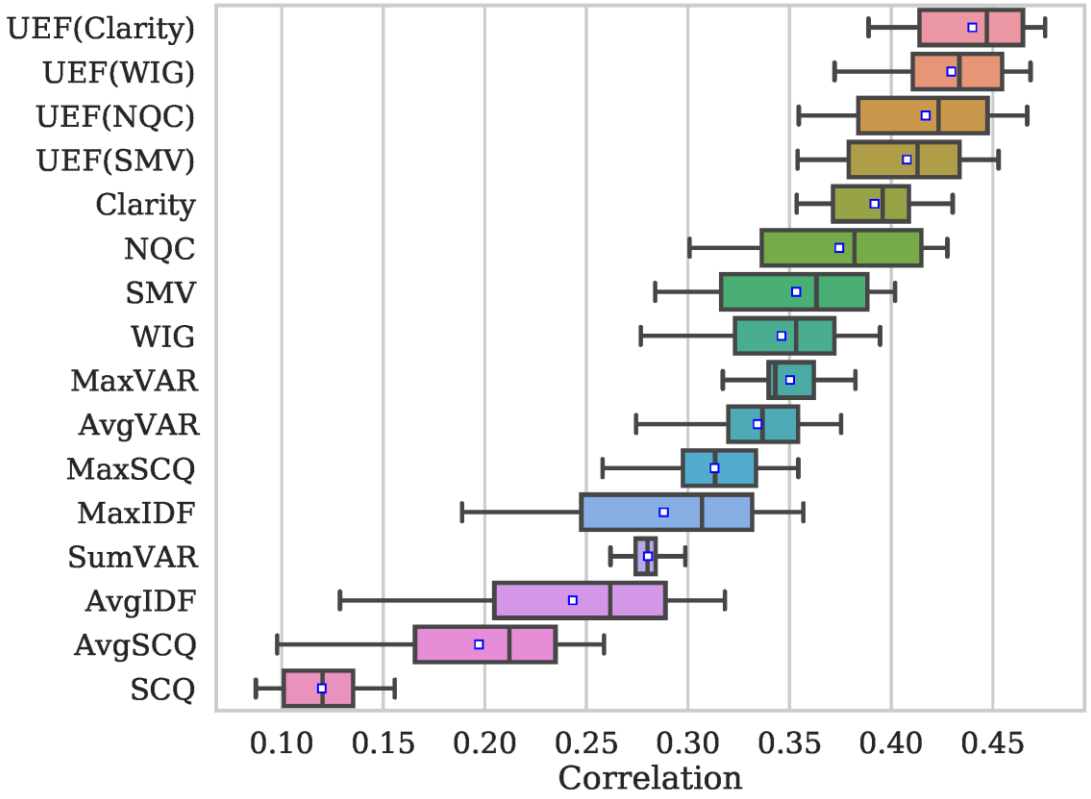
Experimental Setup (Cont.)

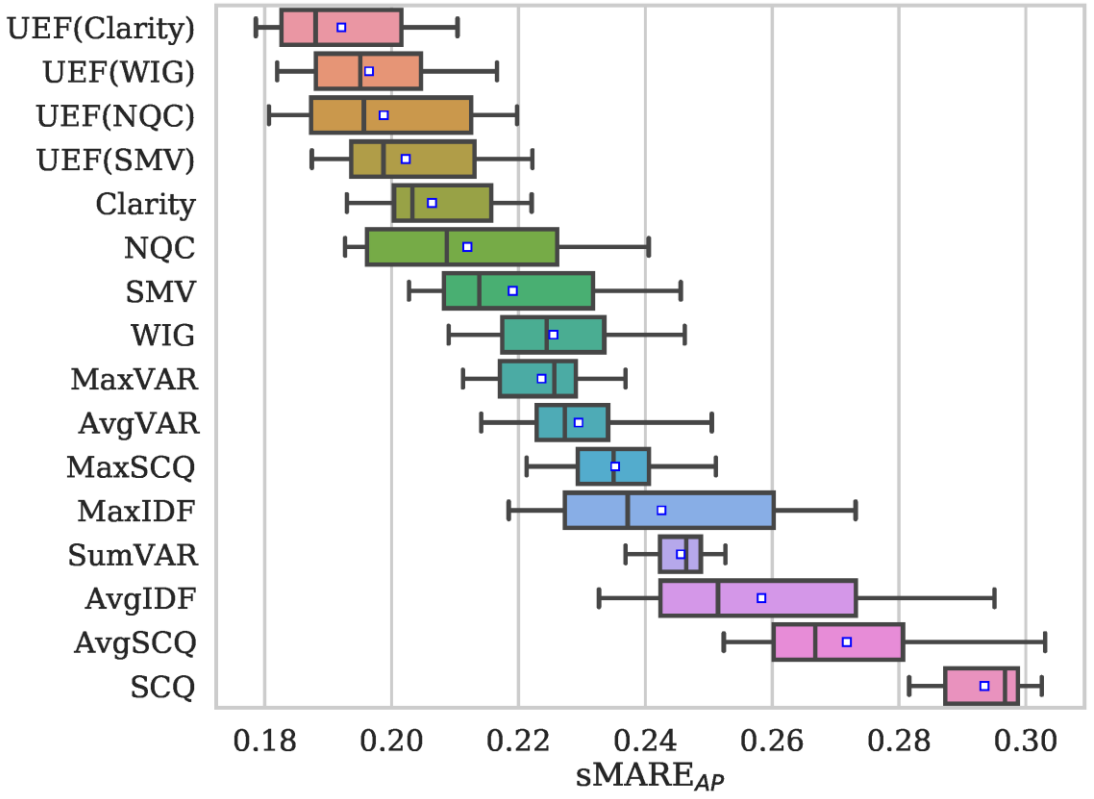
- 3 different stemmer configurations :
  Lovins, *Porter* and no stemming

- 5 different stoplist configurations:
  attire, *zettair*, *indri*, *lingpipe* and no stoplist

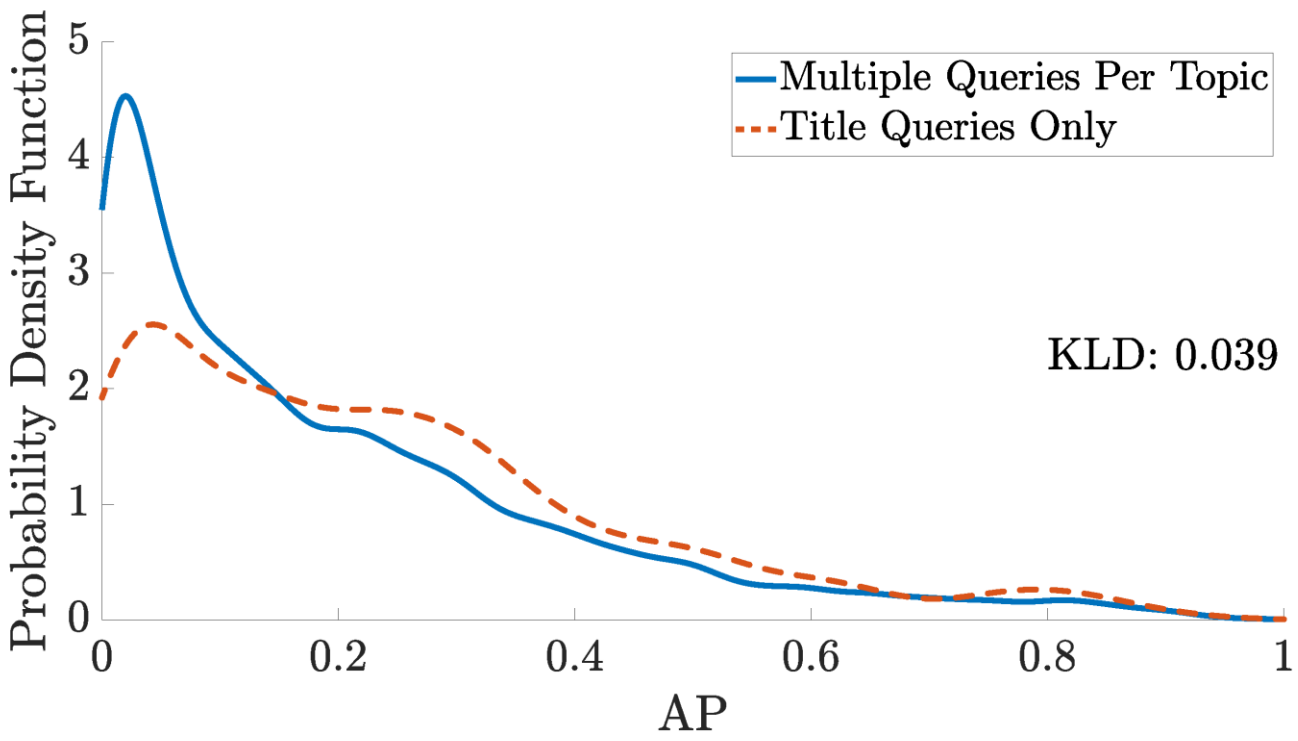
- Total of 15 retrieval pipelines
- Total of 240 QPP-system combinations

Comparison to Traditional QPP Evaluation



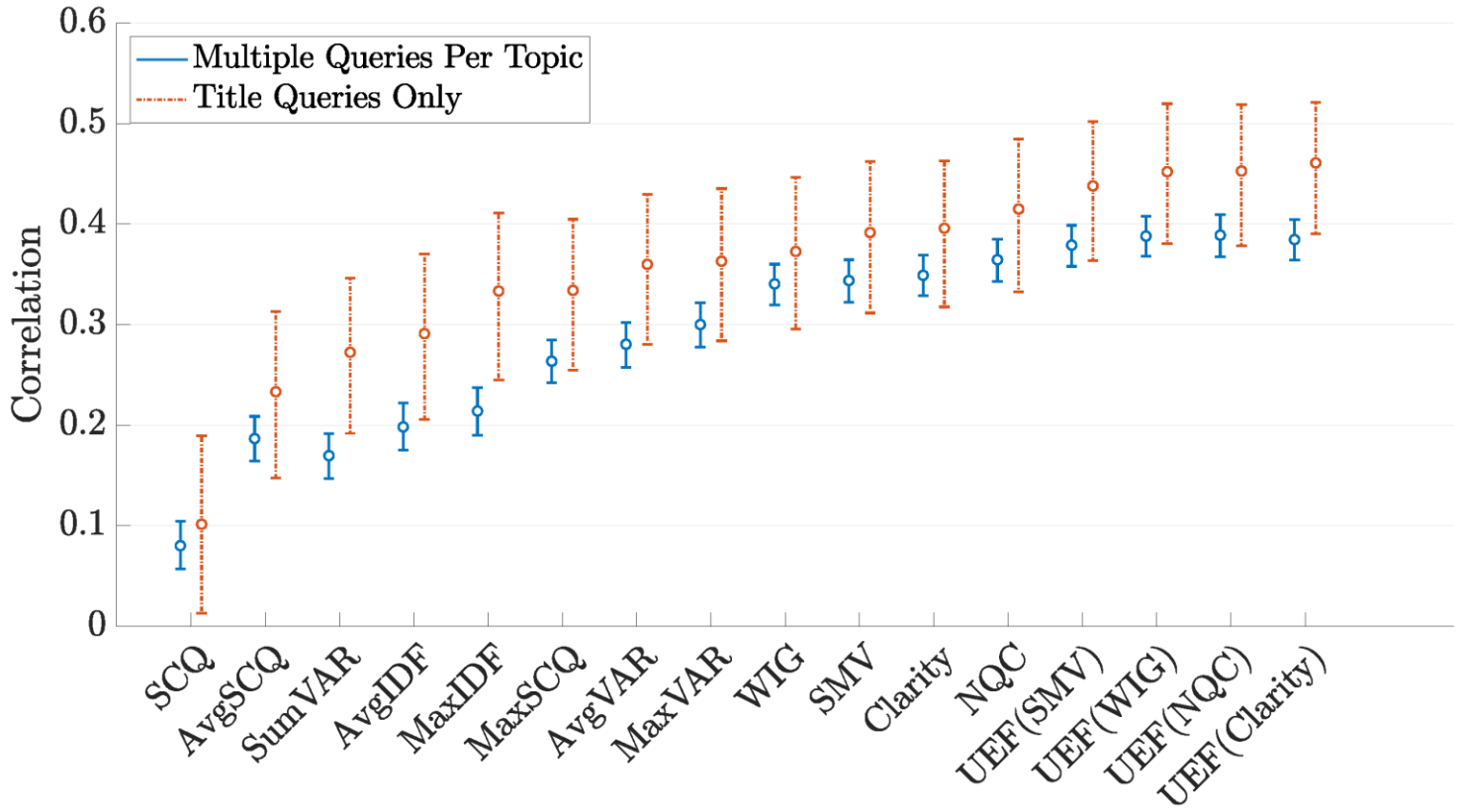


Evaluation with Query Variations



Title queries MAP = 0.25
All queries MAP = 0.21

Correlation Based Comparison



Title queries:

- 57/120 pairs of predictors were found to be statistically significantly different, 47.5%

ANOVA – MD0 Result

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	876.524	248	3.534	168.136	<0.001	0.410
Stoplist	1.185	4	0.296	14.095	<0.001	0.001
Stemmer	5.218	2	2.609	124.108	<0.001	0.004
QPP model	46.569	15	3.105	147.691	<0.001	0.036
Error	1250.538	59490	0.021			
Total	2180.034	59759				

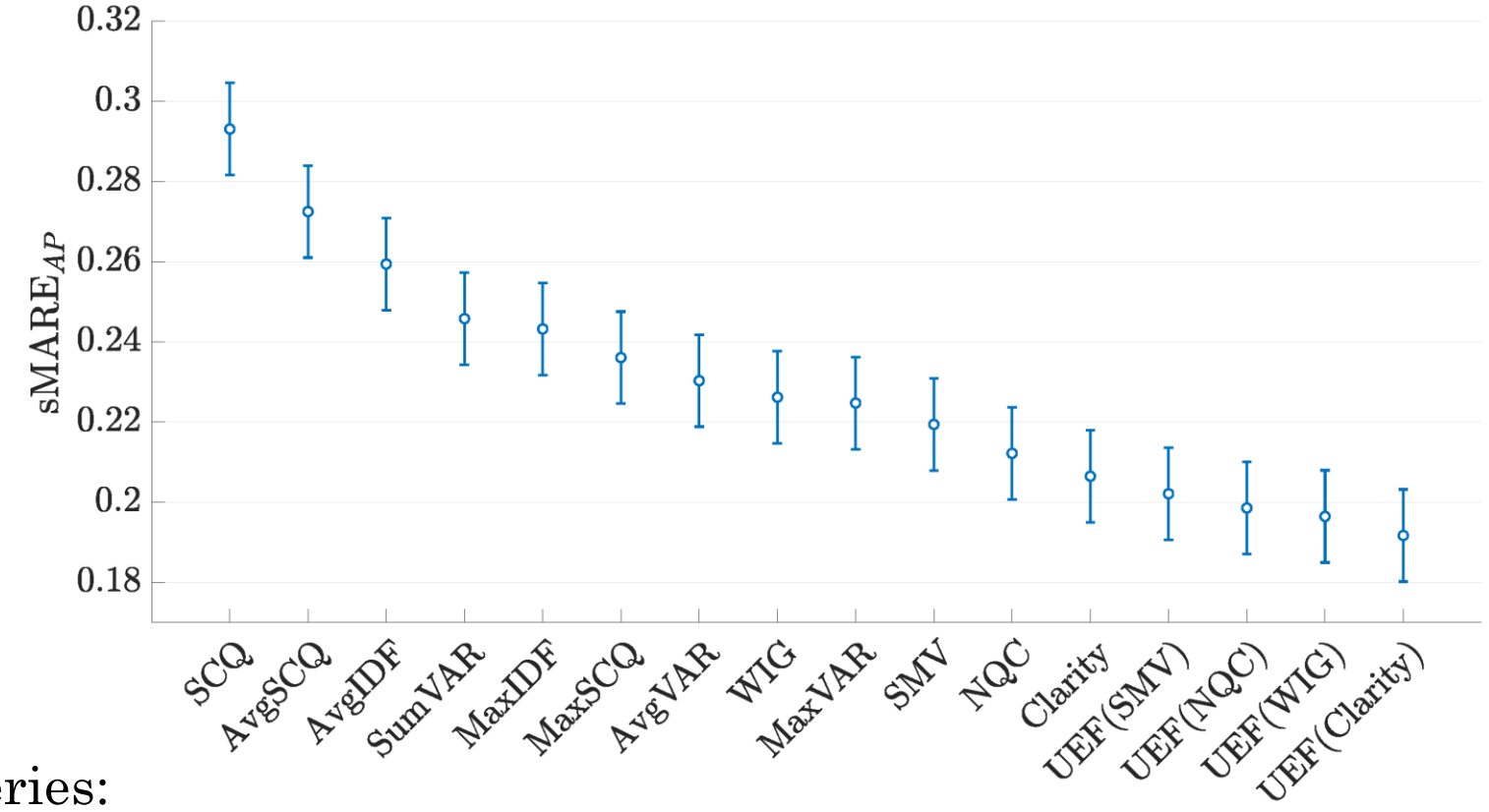
- $\widehat{\omega}^2$ - Strength of Association (SOA), effect size
- Topic has the largest effect on prediction

ANOVA – MD0 Result

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	876.524	248	3.534	168.136	<0.001	0.410
Stoplist	1.185	4	0.296	14.095	<0.001	0.001
Stemmer	5.218	2	2.609	124.108	<0.001	0.004
QPP model	46.569	15	3.105	147.691	<0.001	0.036
Error	1250.538	59490	0.021			
Total	2180.034	59759				

- $\widehat{\omega}^2$ - Strength of Association (SOA), effect size
- Topic has the largest effect on prediction

Tukey's HSD Post-Hoc Analysis



Title queries:

- 96/120 pairs of predictors were found to be statistically significantly different, 80%

Summary

- Defined per topic error
- A distribution of prediction errors
- Enables sound statistical analyses
- Enables factor analysis
- Enables failure analysis

Summary

- Defined per topic error
- A distribution of prediction errors
- Enables sound statistical analyses
- Enables factor analysis
- Enables failure analysis

Summary

- Defined per topic error
- A distribution of prediction errors
- Enables sound statistical analyses
- Enables factor analysis
- Enables failure analysis

Summary

- Defined per topic error
- A distribution of prediction errors
- Enables sound statistical analyses
- Enables factor analysis
- Enables failure analysis

Summary

- Defined per topic error
- A distribution of prediction errors
- Enables sound statistical analyses
- Enables factor analysis
- Enables failure analysis

Questions?

References

- Meng, X.L. et al. 1992. Comparing correlated correlation coefficients. *Psychological Bulletin*. 111, 1 (1992), 172–175.
- Ponte, J.M. and Croft, W.B. 1998. A Language Modeling Approach to Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 1998), 275–281.
- Voorhees, E. and Harman, D.K. 1998. Information Technology: The Sixth Text REtrieval Conference (TREC-6). U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology. Section 3.2.2
- Zhao, Y. et al. 2008. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. ECIR (Berlin, Heidelberg, 2008), 52–64.
- Scholer, F. and Garcia, S. 2009. A Case for Improved Evaluation of Query Difficulty Prediction. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2009), 640–641.
- Hauff, C. et al. 2009. The Combination and Evaluation of Query Performance Prediction Methods. Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (Berlin, Heidelberg, 2009), 301–312.

References (Cont.)

- Benham, R. and Culpepper, J.S. 2017. Risk-reward trade-offs in rank fusion. ACM International Conference Proceeding Series (New York, New York, USA, Dec. 2017), 1–8.
- Thomas, P. et al. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. Proceedings of the 22nd Australasian Document Computing Symposium (New York, NY, USA, 2017).
- Zendel, O. et al. 2019. Information Needs, Queries, and Query Performance Prediction. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2019), 395–404.
- Cronen-Townsend, S. et al. 2004. A Framework for Selective Query Expansion. Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (New York, NY, USA, 2004), 236–237.
- Scholer, F. et al. 2004. Query association surrogates for Web search. Journal of the American Society for Information Science and Technology. 55, 7 (May 2004), 637–650.
- Cronen-Townsend, S. et al. 2002. Predicting Query Performance. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2002), 299–306.

References (Cont.)

- Shtok, A. et al. 2012. Predicting query performance by query-drift estimation. ACM Transactions on Information Systems. 30, 2 (May 2012), 1–35.
- Zhou, Y. and Croft, W.B. 2007. Query Performance Prediction in Web Search Environments. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2007), 543–550.
- Tao, Y. and Wu, S. 2014. Query Performance Prediction By Considering Score Magnitude and Variance Together. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (New York, NY, USA, 2014), 1891–1894.
- Shtok, A. et al. 2010. Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2010), 259–266.
- Webber, W. et al. 2010. A similarity measure for indefinite rankings. ACM Transactions on Information Systems. 28, 4 (Nov. 2010), 1–38.
- Roitman, H. et al. 2019. A Study of Query Performance Prediction for Answer Quality Determination. Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (New York, NY, USA, 2019), 43–46.