

The Combination and Evaluation of Query Performance Prediction Methods

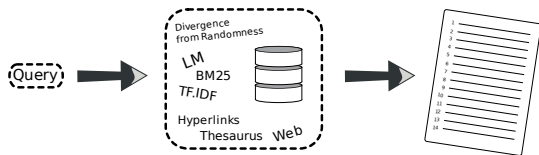
Claudia Hauff¹ Leif Azzopardi² Djoerd Hiemstra¹

¹University of Twente, NL

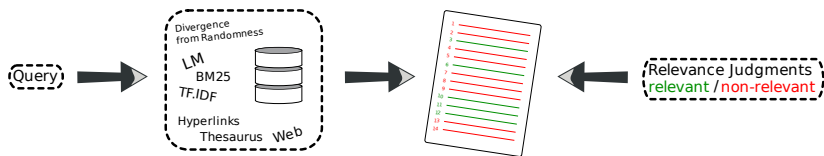
²University of Glasgow, UK

8th April 2009

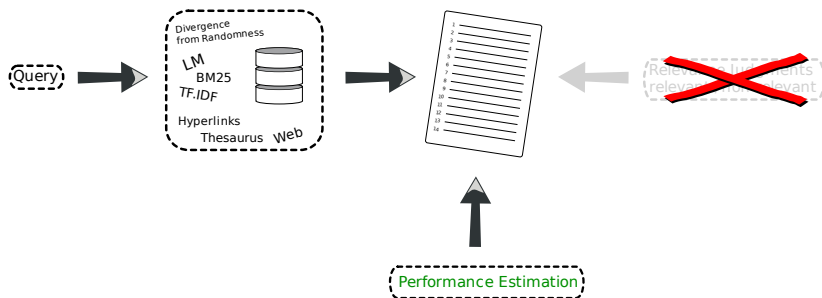
Introduction



Introduction



Introduction



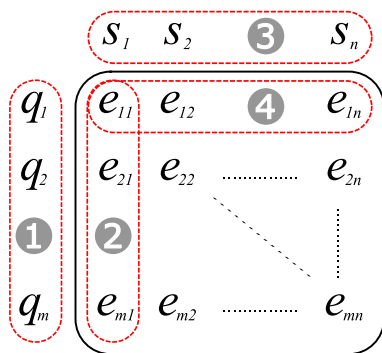
Applications

- TREC-style evaluations without rel. judgments
- Selective query expansion
- Meta-search
- Ask users for query refinement when necessary
- Recognize missing aspects in a collection

Outline

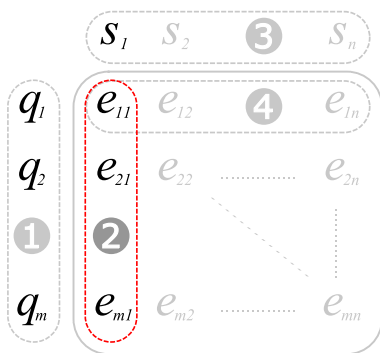
- 1 Introduction & Motivation
- 2 Evaluation Strategies
- 3 Predictors/Combinations
- 4 Experiments
- 5 Conclusions

What can be predicted?



A set of queries $\mathbf{q}_{\{1\dots m\}}$ and a set of retrieval systems $s_{\{1\dots n\}}$ are evaluated on a collection, resulting in a $m \times n$ matrix of retrieval effectiveness values e_{ij} .

In this work ...



A set of queries $\mathbf{q}_{\{1\dots m\}}$ and one retrieval systems s_1 , are evaluated on a collection, resulting in a $m \times 1$ matrix of retrieval effectiveness values e_{i1} .

Task-dependent Evaluation Measures

Given a query q , a collection C , an external source E and a ranking function R

Query Difficulty

$$f_{diff}(q, C, E, R) \rightarrow \{0, 1\}$$

Common measures: [\[Voorhees'03\]](#) [\[Vinay et al.'06\]](#)

Task-dependent Evaluation Measures

Given a query q , a collection C , an external source E and a ranking function R

Query Difficulty

$$f_{diff}(q, C, E, R) \rightarrow \{0, 1\}$$

Common measures: [\[Voorhees'03\]](#) [\[Vinay et al.'06\]](#)

Query Performance

$$f_{perf}(q, C, E, R) \rightarrow \mathbb{R}$$

Common measures: Kendall's τ , Spearman's ρ

Task-dependent Evaluation Measures

Given a query q , a collection C , an external source E and a ranking function R

Query Difficulty

$$f_{diff}(q, C, E, R) \rightarrow \{0, 1\}$$

Common measures: [\[Voorhees'03\]](#) [\[Vinay et al.'06\]](#)

Query Performance

$$f_{perf}(q, C, E, R) \rightarrow \mathbb{R}$$

Common measures: Kendall's τ , Spearman's ρ

Normalized Query Performance

$$f_{norm}(q, C, E, R) \rightarrow [0, 1]$$

Common measure: Linear correlation coefficient r

Evaluation Methodology of f_{norm}

- Given a list of predicted scores \hat{Y} and a list of retrieval effectiveness (e.g. Average Precision) scores Y

$$r = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_{\hat{Y}}\sigma_Y}, \quad r \in [-1, 1]$$

- r indicates the direction/strength of a linear relationship between Y and \hat{Y}
- Common assumption: a higher $|r|$ is sufficient proof of a better algorithm

Evaluation Methodology of f_{norm}

- Given a list of predicted scores \hat{Y} and a list of retrieval effectiveness (e.g. Average Precision) scores Y

$$r = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_{\hat{Y}}\sigma_Y}, \quad r \in [-1, 1]$$

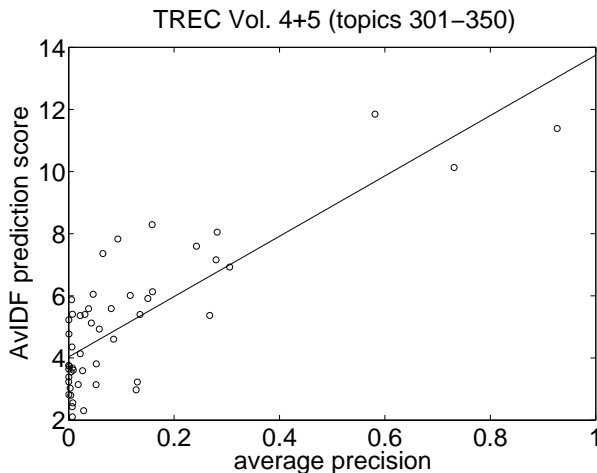
- r indicates the direction/strength of a linear relationship between Y and \hat{Y}
- Common assumption: a higher $|r|$ is sufficient proof of a better algorithm

A More Principled Approach

Report the confidence interval of r and/or test the statistical significance of the *difference* [Meng et al.'92].

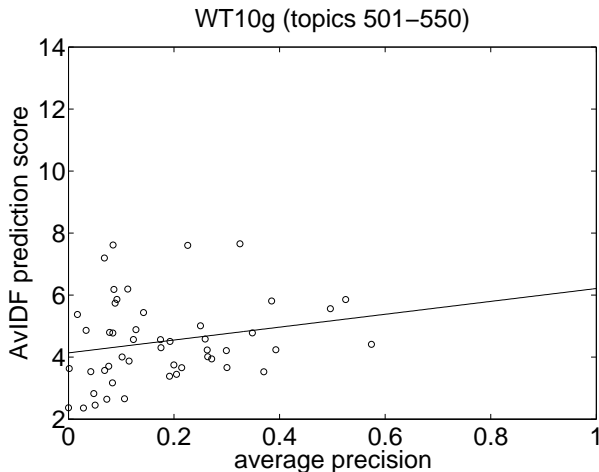
Example I

TF.IDF based retrieval system, $MAP = 0.11$, $r = 0.81$



Example II

Language Modeling based retrieval system, $MAP = 0.18$, $r = 0.22$

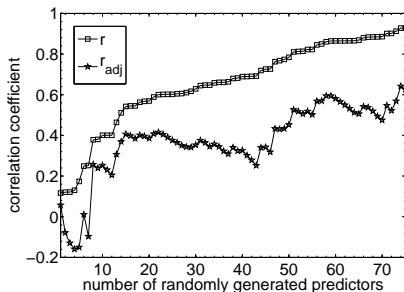


Evaluating Predictor Combinations

- Few attempts to combine (many) predictors
- Reported are either r or r_{adj}^2

Evaluating Predictor Combinations

- Few attempts to combine (many) predictors
- Reported are either r or r_{adj}^2



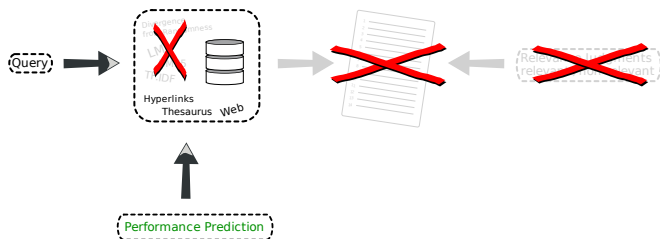
Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{m} \sum_i (y_i - \hat{y}_i)^2}$$

- *RMSE* is easier to interpret
- *RMSE*² is minimized in linear regression, i.e. predictor obtaining highest $|r|$ also obtains lowest *RMSE*
- To avoid mixing training and test data, multiple linear regression experiments are evaluated with leave-one-out cross-validation

Pre-retrieval Predictors

Pre-retrieval predictors predict the performance of a query without considering the ranked list of results.



Heuristics

Specificity

A query performs better with increased specificity.

Heuristics

Specificity

A query performs better with increased specificity.

Ambiguity

The fewer senses/contexts the query terms appear in, the better.

Heuristics

Specificity

A query performs better with increased specificity.

Ambiguity

The fewer senses/contexts the query terms appear in, the better.

Term relatedness

The more the query terms are related, the better the query.

Heuristics

Specificity

A query performs better with increased specificity.

Ambiguity

The fewer senses/contexts the query terms appear in, the better.

Term relatedness

The more the query terms are related, the better the query.

Ranking sensitivity

A query performs well, if the retrieval system can rank the query-terms containing documents.

Examples of Predictors

- Average Query Term Length (*AvQL*) [*Mothe & Tanguy'05*]
- Max. Inverse Document Frequency (*MaxIDF*) [*Scholer et al.'04*]
- Average Set Coherence (*AvQC*) [*He et al.'08*]
- Average Pointwise Mutual Information (*AvPMI*)
- Max. Term Weight Variability (*MaxVAR*) [*Zhao et al.'08*]

Penalized Regression I

- Multiple Linear Regression (OLS):

$$\min_{\beta} \|Y - X\beta\|_2^2$$

- prone to overfitting
- lack of model interpretation

$Y_{m \times 1}$	AP scores
$X_{m \times p}$	predictors
$\beta_{p \times 1}$	predictor coefficients

- In *microarray data analysis*, penalized regression approaches are used to constrain β
 - automatic model selection and shrinkage
 - accurate predictions
 - interpretable models

Penalized Regression II

- Least Abs. Shrinkage and Selection Operator [*Tibshirani'96*]
 - $\min_{\beta} \|Y - X\beta\|_2^2 + \theta \sum_{j=1}^p |\beta_j|$
 - of similar predictors, only one tends to be included in the model
- Elastic Net [*Zhou & Hastie'05*]
 - $\min_{\beta} \|Y - X\beta\|_2^2 + \theta_1 \sum_{j=1}^p |\beta_j| + \theta_2 \sum_{j=1}^p \beta_j^2$
 - highly correlated predictors are brought into the model together
- Least Angle Regression [*Efron et al.'04*] to compute the full regularization path (β coefficient matrix of size $p \times p$)
- Selecting β vector
 - Traps: injection of randomly generated predictors
 - Cross-validation
 - Bootstrapped samples

Experimental Setup

	TREC Vol. 4+5	WT10g	GOV2
<i>#documents</i>	528155	1692095	25199132
<i>#unique terms</i>	764376	7081712	32933168
<i>av. doc. length</i>	266.4	377.6	665.3
<i>topics</i>	301-450	451-550	701-850
<i>av. topic length</i>	2.48	2.63	2.97

- Preprocessing: Krovetz stemming & stopword removal
- Retrieval approach: Language Modeling with Dirichlet smoothing ($\mu = 1000$)
- 19 predictors

Results: Statistical Significance

	TREC Vol. 4+5			WT10g			GOV2		
	r_{train}	CI_{train}	$RMSE$	r_{train}	CI_{train}	$RMSE$	r_{train}	CI_{train}	$RMSE$
<i>AvPMI</i>	0.35	[0.21,0.48]	0.207	<u>0.28</u>	[0.09,0.46]	0.191	<u>0.28</u>	[0.12,0.42]	0.187
<i>AvQC</i>	<u>0.46</u>	[0.32,0.58]	0.191	0.16	[-0.04,0.35]	0.196	<u>0.30</u>	[0.14,0.44]	0.184
<i>AvQL</i>	0.13	[-0.03,0.29]	0.215	-0.14	[-0.32,0.07]	0.197	0.01	[-0.15,0.17]	0.194
<i>MaxIDF</i>	<u>0.53</u>	[0.41,0.64]	0.181	<u>0.29</u>	[0.10,0.46]	0.187	<u>0.33</u>	[0.18,0.47]	0.181
<i>MaxVAR</i>	<u>0.51</u>	[0.38,0.62]	0.182	<u>0.41</u>	[0.23,0.56]	0.184	<u>0.41</u>	[0.27,0.54]	0.176

Results: Statistical Significance

	TREC Vol. 4+5			WT10g			GOV2		
	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE
<i>AvPMI</i>	0.35	[0.21,0.48]	0.207	<u>0.28</u>	[0.09,0.46]	0.191	<u>0.28</u>	[0.12,0.42]	0.187
<i>AvQC</i>	<u>0.46</u>	[0.32,0.58]	0.191	0.16	[-0.04,0.35]	0.196	<u>0.30</u>	[0.14,0.44]	0.184
<i>AvQL</i>	0.13	[-0.03,0.29]	0.215	-0.14	[-0.32,0.07]	0.197	0.01	[-0.15,0.17]	0.194
<i>MaxIDF</i>	<u>0.53</u>	[0.41,0.64]	0.181	<u>0.29</u>	[0.10,0.46]	0.187	<u>0.33</u>	[0.18,0.47]	0.181
<i>MaxVAR</i>	<u>0.51</u>	[0.38,0.62]	0.182	<u>0.41</u>	[0.23,0.56]	0.184	<u>0.41</u>	[0.27,0.54]	0.176

Results: Statistical Significance

	TREC Vol. 4+5			WT10g			GOV2		
	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE
<i>AvPMI</i>	0.35	[0.21,0.48]	0.207	<u>0.28</u>	[0.09,0.46]	0.191	<u>0.28</u>	[0.12,0.42]	0.187
<i>AvQC</i>	<u>0.46</u>	[0.32,0.58]	0.191	0.16	[-0.04,0.35]	0.196	<u>0.30</u>	[0.14,0.44]	0.184
<i>AvQL</i>	0.13	[-0.03,0.29]	0.215	-0.14	[-0.32,0.07]	0.197	0.01	[-0.15,0.17]	0.194
<i>MaxIDF</i>	<u>0.53</u>	[0.41,0.64]	0.181	<u>0.29</u>	[0.10,0.46]	0.187	<u>0.33</u>	[0.18,0.47]	0.181
<i>MaxVAR</i>	<u>0.51</u>	[0.38,0.62]	0.182	<u>0.41</u>	[0.23,0.56]	0.184	<u>0.41</u>	[0.27,0.54]	0.176

Results: Statistical Significance

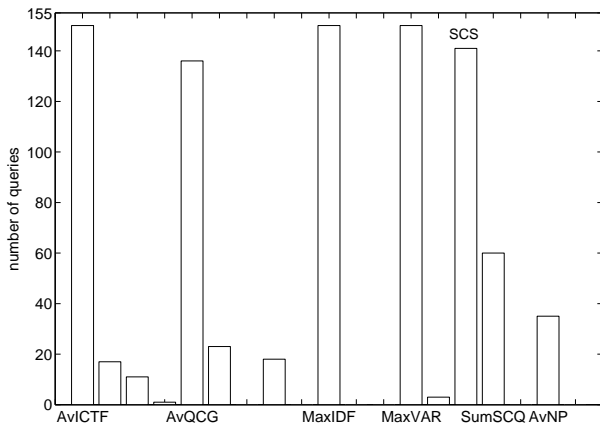
	TREC Vol. 4+5			WT10g			GOV2		
	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE
<i>AvPMI</i>	0.35	[0.21,0.48]	0.207	<u>0.28</u>	[0.09,0.46]	0.191	<u>0.28</u>	[0.12,0.42]	0.187
<i>AvQC</i>	<u>0.46</u>	[0.32,0.58]	0.191	0.16	[-0.04,0.35]	0.196	<u>0.30</u>	[0.14,0.44]	0.184
<i>AvQL</i>	0.13	[-0.03,0.29]	0.215	-0.14	[-0.32,0.07]	0.197	0.01	[-0.15,0.17]	0.194
<i>MaxIDF</i>	<u>0.53</u>	[0.41,0.64]	0.181	<u>0.29</u>	[0.10,0.46]	0.187	<u>0.33</u>	[0.18,0.47]	0.181
<i>MaxVAR</i>	<u>0.51</u>	[0.38,0.62]	0.182	<u>0.41</u>	[0.23,0.56]	0.184	<u>0.41</u>	[0.27,0.54]	0.176

Results: Penalized regression

In bold, improvements over the best single predictor per collection are shown.

Predictor	TREC Vol. 4+5			WT10g			GOV2		
	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE	r_{train}	CI_{train}	RMSE
<i>OLS</i>	0.69	[0.60,0.77]	0.188	0.64	[0.51,0.74]	0.208	0.52	[0.39,0.63]	0.190
<i>LARS-Traps</i>	0.59	[0.47,0.68]	0.179	0.52	[0.36,0.65]	0.187	0.44	[0.30,0.56]	0.178
<i>LARS-CV</i>	0.68	[0.59,0.76]	0.183	0.53	[0.38,0.66]	0.178	0.46	[0.33,0.58]	0.184
<i>BOLASSO</i>	0.59	[0.47,0.68]	0.181	0.43	[0.25,0.58]	0.198	0.43	[0.28,0.55]	0.180
<i>Elastic Net</i>	0.69	[0.60,0.77]	0.182	0.52	[0.35,0.65]	0.182	0.46	[0.32,0.57]	0.178

Model selection: LARS-Traps (TREC 4+5)



Conclusions & Future Work

- Conclusions
 - Predictor performance is highly collection dependent
 - A number of predictors are not significantly different from the best performing predictor
 - Predictor combinations show improvements, though not stable across all collections
- Future Work
 - Evaluation in context of a particular application
 - Combinations of post-retrieval predictors