

Frameworks for Evaluating and Meta-Evaluating Search Engines

Ph.D Thesis Defense

Peter B. Golbus

Advisor: Javed A. Aslam

Northeastern University
College of Computer and Information Science

June 19th, 2014

Evaluation: The Bottleneck in IR Research

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Web search, Diversification, Topic Distillation, Knowledge Acquisition, Temporal Summarization, Legal search, Enterprise search, Microblog search...
- Little to no formal problem description

Evaluation: The Bottleneck in IR Research

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Web search, Diversification, Topic Distillation, Knowledge Acquisition, Temporal Summarization, Legal search, Enterprise search, Microblog search...
- Little to no formal problem description
- We don't really know what we're trying to do, so...

Evaluation: The Bottleneck in IR Research

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Web search, Diversification, Topic Distillation, Knowledge Acquisition, Temporal Summarization, Legal search, Enterprise search, Microblog search...
- Little to no formal problem description
- We don't really know what we're trying to do, so...
 - 1 Are we improving the right thing?
 - 2 How do we know when we've done it?

Outline

Meta-Evaluating Search Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation
- 4 Information Difference
- 5 Conclusion

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation
- 4 Information Difference
- 5 Conclusion

The Ad Hoc Retrieval Task

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Information need expressed in detailed narrative e.g. Documents discussing electrical output of the Three Gorges Dam
 - Known as topics or queries
- Find documents containing even one pertinent fact
 - Set of “relevance labels” known as a QREL
- Recall-oriented: user interacts with *all* ranked documents

The Ad Hoc Retrieval Task

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Information need expressed in detailed narrative e.g. Documents discussing electrical output of the Three Gorges Dam
 - Known as topics or queries
 - Find documents containing even one pertinent fact
 - Set of “relevance labels” known as a QREL
 - Recall-oriented: user interacts with *all* ranked documents
- e.g. Annual NIST-sponsored Text REtrieval Conference (TREC)

Ad Hoc Evaluation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

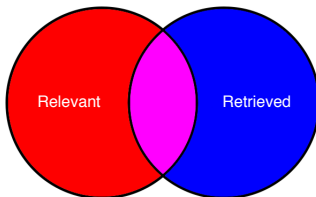
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- Precision: percentage of retrieved documents that are relevant

$$Prec@k = \frac{|Relevant| \cap |Retrieved|}{|Retrieved|}$$

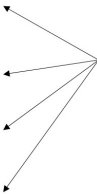
- Recall: percentage of relevant documents that are retrieved

$$Rec@k = \frac{|Relevant| \cap |Retrieved|}{|Relevant|}$$

Ad Hoc Evaluation

- Precision: percentage of retrieved documents that are relevant
- Recall: percentage of relevant documents that are retrieved
- Average Precision: average precision at ranks of relevant documents

Doc	Precision
R	1
N	-
R	0.66
N	-
R	0.60
...	
R	0.27

$$AP = \frac{\sum_{k=1}^{\infty} rel_k \cdot Prec@k}{|relevant|}$$


Web Search

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier

Metasearch

Conclusion

- Information need expressed in 2 or 3 word query
- Find *high quality, highly relevant* documents
- Precision-oriented: user interacts with few documents

Web Search

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier

Metasearch

Conclusion

- Information need expressed in 2 or 3 word query
- Find *high quality, highly relevant* documents
- Precision-oriented: user interacts with few documents
 - Usually

Web Evaluation

- Find *high quality, highly relevant* documents
 - Use graded relevance labels
- Precision-oriented: user interacts with few documents
 - Rank-based discounting
- $eval = \langle \mathbf{gain}, \mathbf{discount} \rangle$

Rank	Grade	Gain	Discount
1	4	15	1
2	0	0	0.5
3	2	3	0.38
4	0	0	0.33
5	3	7	0.30

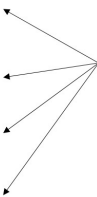
$$(n)DCG = \frac{2^{grade} - 1}{\log(rank + 1)}$$

$$nDCG@5 = 18.24$$

Ad Hoc Evaluation

- Precision: percentage of retrieved documents that are relevant
- Recall: percentage of relevant documents that are retrieved
- Average Precision: average precision at ranks of relevant documents

Doc	Precision
R	1
N	-
R	0.66
N	-
R	0.60
...	
R	0.27

$$AP = \frac{\sum_{k=1}^{\infty} rel_k \cdot Prec@k}{|relevant|}$$


Web Evaluation

- Find *high quality, highly relevant* documents
 - Use graded relevance labels
- Precision-oriented: user interacts with few documents
 - Rank-based discounting
- $eval = \langle \mathbf{gain}, \mathbf{discount} \rangle$

Rank	Grade	Gain	Discount
1	4	15	1
2	0	0	0.5
3	2	3	0.38
4	0	0	0.33
5	3	7	0.30

$$(n)DCG = \frac{2^{grade} - 1}{\log(rank + 1)}$$

$$nDCG@5 = 18.24$$

Web Evaluation

■ $eval = \langle \text{gain}, \text{discount} \rangle$

Rank	Grade	Gain	Discount
1	4	15	1
2	0	0	0.5
3	2	3	0.38
4	0	0	0.33
5	3	7	0.30

$$(n)DCG = \frac{2^{grade} - 1}{\log(rank + 1)}$$

nDCG@5 = 18.24

Doesn't Match Actual Behavior!*

*Golbus *et al.* WWW '14

Goals

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Use targeted meta-evaluations to better understand IR problems
- 2 New, powerful, highly interpretable information-theoretic toolkit
 - Evaluation and meta-evaluation within a single unified framework

- 1 Targeted Meta-Evaluation for Understanding Diversity
 - Document Selection Sensitivity
- 2 Probabilistic Framework for Evaluation and Rank Correlation
 - Information-Theoretic Evaluation Measure
 - Conditional Rank Correlation
- 3 Information Difference
 - Similarity Classifier
 - Selecting Systems for Metasearch

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation
- 4 Information Difference
- 5 Conclusion

Novelty & Diversity

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Subtopics

What is the information need of the query “freddie mercury”

- biography, memorabilia, music?

Goal

Cover as many subtopics as possible by rank k

Two Factors

- 1 Diversity: how well did you order the documents?

Two Factors

- 1 Diversity: how well did you order the documents?
- 2 Performance: how many good documents did you find?

Evaluation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Diversity measures are traditional measures with an added diversity component
 - e.g. Intent Aware measures: evaluate with respect to each subtopic and compute the average
 - nDCG $\rightarrow$ nDCG-IA
- Are we really measuring diversity?

Evaluation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Diversity measures are traditional measures with an added diversity component
 - e.g. Intent Aware measures: evaluate with respect to each subtopic and compute the average
 - nDCG $\rightarrow$ nDCG-IA
- Are we really measuring diversity?

No!

Measure Sensitivity

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

How **sensitive** to changes in system performance?

- Discriminative power: percentage of system pairs that are statistically significantly different

Measure Sensitivity

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

How **sensitive** to changes in system performance?

- Discriminative power: percentage of system pairs that are statistically significantly different
- Sensitivity is necessary but not sufficient

Measure Sensitivity

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

How **sensitive** to changes in system performance?

- Discriminative power: percentage of system pairs that are statistically significantly different
- Sensitivity is necessary but not sufficient

What is the measure sensitive *to*?

Does Document Order Matter?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

ERR-IA	2010	2011
Actual	0.571	0.544

D$\#$-nDCG	2010	2011
Actual	0.583	0.600

Table: Discriminative power on actual runs

Does Document Order Matter?

ERR-IA	2010	2011
Actual	0.571	0.544
Permutations	0.039	0.031

D_#-nDCG	2010	2011
Actual	0.583	0.600
Permutations	0.036	0.019

Table: Discriminative power on actual runs and permutations of relevant documents

What Went Wrong?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Two Factors

- 1 Diversity: how well did you order the documents?
- 2 Performance: how many good documents did you find?

Example

- biography
- memorabilia
- music
- Not relevant

What Went Wrong?

Two Factors

- 1 Diversity: how well did you order the documents?
- 2 Performance: how many good documents did you find?

Bad performance $\Rightarrow$ bad diversity



What Went Wrong?

Two Factors

- 1 Diversity: how well did you order the documents?
- 2 Performance: how many good documents did you find?

Good diversity $\Rightarrow$ good performance



What Went Wrong?

Two Factors

- 1 Diversity: how well did you order the documents?
- 2 Performance: how many good documents did you find?

Good performance $\nRightarrow$ good diversity



What Went Wrong?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Two Factors

- 1 Diversity: how well did you order the documents?
 - This does not matter
- 2 Performance: how many good documents did you find?
 - This does

Research Bottleneck Example

The Problem

We don't really know what we're trying to do, so...

- 1 Are we improving the right thing?
- 2 How do we know when we've done it?

Recent Work

- Hypothesis: Metasearch improves diversity
 - Metasearch improves performance
 - Improving performance improves diversity *evaluation scores*

Research Bottleneck Example

The Problem

We don't really know what we're trying to do, so...

- 1 Are we improving the right thing?
- 2 How do we know when we've done it?

Recent Work

- Hypothesis: Metasearch improves diversity
 - Metasearch improves performance
 - Improving performance improves diversity *evaluation scores*

Hypothesis not tested by standard diversity evaluation paradigm!

Document Selection Sensitivity

- How do we control for performance?
 - Use only relevant documents
- What do we want to know?
 - Does document order impact evaluation score?

Document Selection Sensitivity (dss)

- For a measure M , evaluate random permutations of relevant documents
- Document Selection Sensitivity is the *coefficient of variation*

$$dss(M) = \frac{\text{standard deviation}}{\text{mean}}$$

Document Selection Sensitivity

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

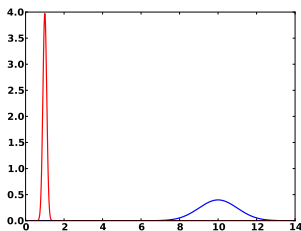
Framework

RIC
Conditional
Rank Correlation

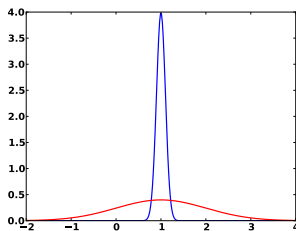
ID

Classifier
Metasearch

Conclusion



Same



Different

- Scale-invariant measure of dispersion
- A DSS of .1 means 68% of scores are within $\pm 10\%$ of the average score (assuming normally distributed)

Improved Measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

DSS	ERR	DCG	RBP
Cascade	0.204	0.152	0.170
D#	0.116	0.096	0.111
$\alpha\#$ -IA	0.871	0.801	0.801

- Intuition: Focus on difficult topics
- *i.e.* leverage as much diversity information from the collection as you can

Targeted Meta-Evaluation for Understanding Diversity

- Document Selection Sensitivity

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation**
 - Information-Theoretic Evaluation Measure
 - Conditional Rank Correlation
- 4 Information Difference
- 5 Conclusion

Observation 1: From Geometric to Probabilistic

- Currently: Vector space model of evaluation

$$\text{eval}(S) = \langle \mathbf{gain}, \mathbf{discount} \rangle$$

- Desired: Probabilistic model

$$\text{eval}(S) = I(QREL; RANKED_LIST)$$

Observation 2: Comparing Orderings

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

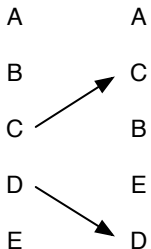
Framework

RIC
Conditional
Rank Correlation

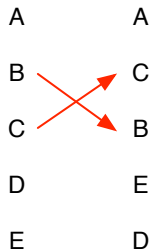
ID

Classifier
Metasearch

Conclusion



Concordant



Discordant

Kendall's τ

- $c = \#$ of concordant pairs, $d = \#$ of discordant pairs
- $\tau = \frac{c-d}{c+d} \in [-1, 1]$

Observation 2: Comparing Orderings

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Kendall's τ depends on preferences, not ranks
- QRELs and ranked lists encode document preferences
- Lets measure “rank” correlation between these preferences

Encoding Preferences

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

{Highly Relevant Docs}
 {Relevant Docs}
 {Non-Relevant Docs}
 {Unjudged Docs}

QREL

Doc 1
Doc 2
⋮
Doc k
{Unretrieved Docs}

Ranked List

Probabilistic Framework

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Sample space
- 2 Distribution over sample space
- 3 Random variables

Probabilistic Framework

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Sample space
 - $\Omega = \text{all } 2 \cdot \binom{n}{2} \text{ document pairs}$
- 2 Distribution over sample space
- 3 Random variables

Probabilistic Framework

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Sample space
 - $\Omega = \text{all } 2 \cdot \binom{n}{2} \text{ document pairs}$
- 2 Distribution over sample space
 - $P = U$
- 3 Random variables

Probabilistic Framework

1 Sample space

- $\Omega = \text{all } 2 \cdot \binom{n}{2} \text{ document pairs}$

2 Distribution over sample space

- $P = U$

3 Random variables

- For a ranked list R , $X_R: \Omega \rightarrow \{-1, +1\}$

$$X_R [(d_i, d_j)] = \begin{cases} 1 & \text{if } d_i \text{ appears before } d_j \text{ in } R. \\ -1 & \text{otherwise.} \end{cases}$$

- $E[X_R] = 0$

Probabilistic Framework

1 Sample space

- $\Omega = \text{all } 2 \cdot \binom{n}{2} \text{ document pairs}$

2 Distribution over sample space

- $P = U$

3 Random variables

- For a ranked list R , $X_R: \Omega \rightarrow \{-1, +1\}$

$$X_R[(d_i, d_j)] = \begin{cases} 1 & \text{if } d_i \text{ appears before } d_j \text{ in } R. \\ -1 & \text{otherwise.} \end{cases}$$

- $E[X_R] = 0$

For two lists R and S

- $E[X_R \cdot X_S] = \frac{2c-2d}{2(c+d)} = \tau(R, S)$

Power of Framework

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Power of Framework

- 1 Generalizes to partial orderings
- 2 Flexibility of random variables
- 3 Information-theoretic interpretation

Information Theory

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

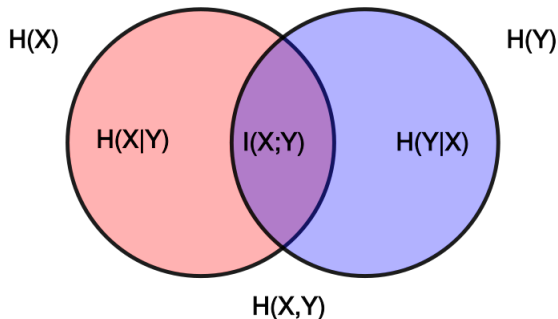
Framework

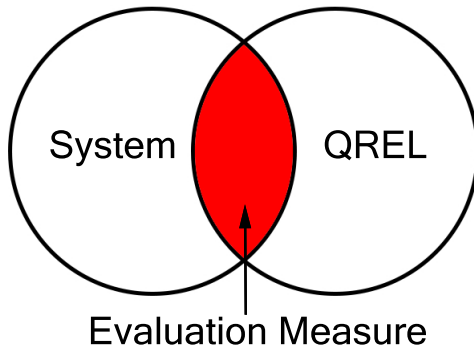
RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion





$$I(\text{System}; \text{QREL})$$

Conditional Rank Correlation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

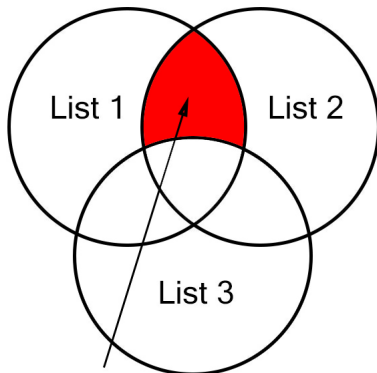
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Conditional Rank
Correlation

$$I(List_1; List_2 \mid List_3)$$

Conditional Rank Correlation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

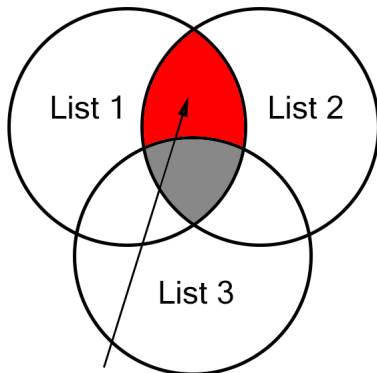
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Conditional Rank
Correlation

$$I(List_1; List_2 | List_3)$$

Information Difference

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

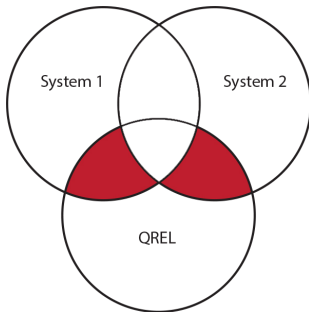
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



$$id(S_1, S_2) = I(S_1; Q \mid S_2) + I(S_2; Q \mid S_1)$$

Information Difference

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

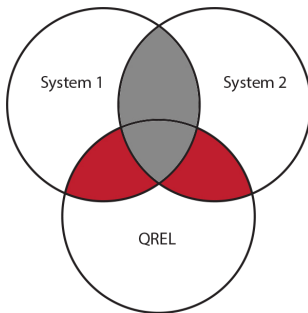
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



$$id(S_1, S_2) = I(S_1; Q \mid S_2) + I(S_2; Q \mid S_1)$$

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

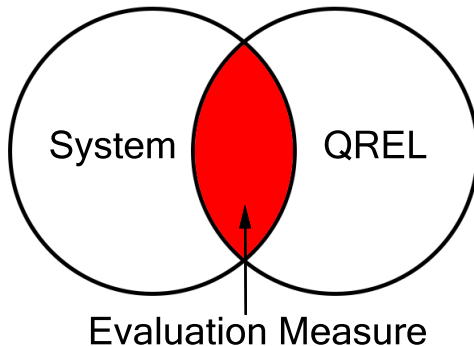
2 Targeted Meta-Evaluation for Understanding Diversity

3 Probabilistic Framework for Evaluation and Rank Correlation

- Information-Theoretic Evaluation Measure
- Conditional Rank Correlation

4 Information Difference

5 Conclusion



$$I(\text{System}; \text{QREL})$$

Evaluation Measure

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Relevance Information Correlation

$$RIC(S) = I(R_S; Q)$$

Evaluation Measure

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Relevance Information Correlation

$$RIC(S) = I(R_S; Q)$$

$RIC(S_1, \dots, S_n)$ is well-defined

Evaluation Measure

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Relevance Information Correlation

$$RIC(S) = I(R_S; Q)$$

$RIC(S_1, \dots, S_n)$ is well-defined

- Upper bound on metasearch

Probabilistic Framework

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Sample space
- 2 Distribution over sample space
 - $RIC@k^*$
- 3 Random variables

*Golbus & Aslam. SIGIR '14.

“Rank” of a Relevance Grade

Consider ranked lists consistent with the QREL:

Highly Relevant

⋮

Highly Relevant

Relevant

⋮

Relevant

Non-Relevant

⋮

Non-Relevant

- R_g := number of documents with relevance grade g
- k_{min} := minimum rank of a document with grade g
- k_{max} := maximum rank of a document with grade g

Probability of a Document

- A rank discount function can be interpreted as a probability distribution.*

- $P_{DCG}(k) = \frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)}$

- Probability of a document

$$\begin{aligned} P(doc) &= \alpha \frac{\sum_{i=k_{min}}^{k_{max}} \frac{1}{\log_2(i+1)} - \frac{1}{\log_2(i+2)}}{R_g} \\ &= \alpha \frac{\frac{1}{\log_2(k_{min}+1)} - \frac{1}{\log_2(k_{max}+2)}}{R_g} \end{aligned}$$

- Probability of a document pair
 - $P(d_i, d_j) = \beta P(d_i)P(d_j)$

*Carterette. SIGIR '11.

RIC vs nDCG @ 20

Meta-Evaluating Search Engines

Golbus

Motivation

Diversity

Framework

RIC

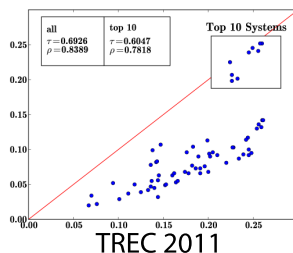
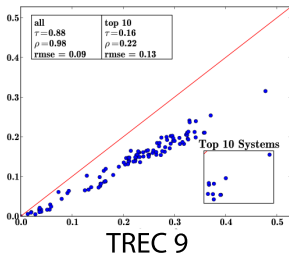
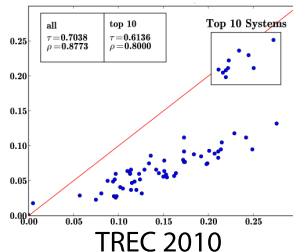
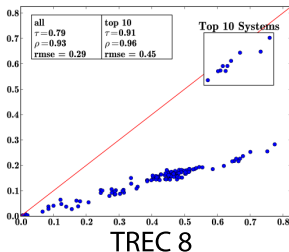
Conditional Rank Correlation

ID

Classifier

Metasearch

Conclusion



Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

2 Targeted Meta-Evaluation for Understanding Diversity

3 Probabilistic Framework for Evaluation and Rank Correlation

- Information-Theoretic Evaluation Measure
- Conditional Rank Correlation

4 Information Difference

5 Conclusion

Conditional Rank Correlation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

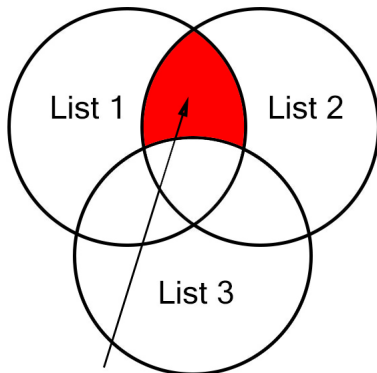
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Conditional Rank
Correlation

$$I(List_1; List_2 \mid List_3)$$

Diversity Dominated by Performance

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Diversity measures are highly rank-correlated
- Diversity evaluation is dominated by performance
- Hypothesis: Correlation due to performance, not diversity

Diversity Dominated by Performance

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Diversity measures are highly rank-correlated
- Diversity evaluation is dominated by performance
- Hypothesis: Correlation due to performance, not diversity

How do we test this?

Rank Correlation

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

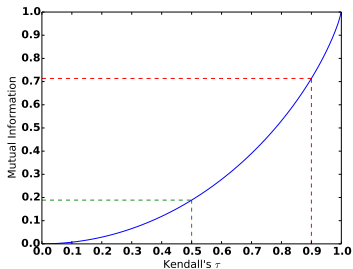
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Rank Correlation

$$\blacksquare I(X_R; X_S) = \frac{1+\tau}{2} \log(1+\tau) + \frac{1-\tau}{2} \log(1-\tau)$$

Conditional Rank Correlation

Meta-Evaluating Search Engines

Golbus

Motivation

Diversity

Framework

RIC

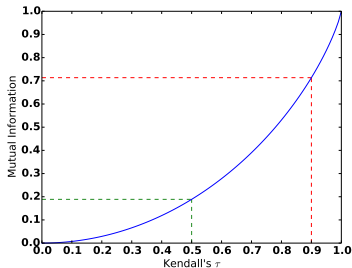
Conditional Rank Correlation

ID

Classifier

Metasearch

Conclusion



Conditional Rank Correlation

- $\tau_I(R, S) \stackrel{\text{def}}{=} I(X_R; X_S)$
- $\tau_I(R, S \mid T) \stackrel{\text{def}}{=} I(X_R; X_S \mid X_T)$

Comparing Evaluation Measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier

Metasearch

Conclusion

	Measure 1	Measure 2
1.	Bing	Google
2.	Yandex	Bing
3.	Google	Yandex
4.	Baidu	Baidu

Comparing Evaluation Measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC

Conditional
Rank Correlation

ID

Classifier

Metasearch

Conclusion

	Measure 1	Measure 2
1.	Bing	Google
2.	Yandex	Bing
3.	Google	Yandex
4.	Baidu	Baidu

Concordant vs. Discordant

Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

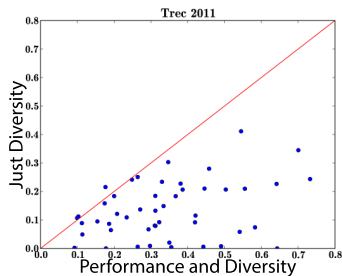
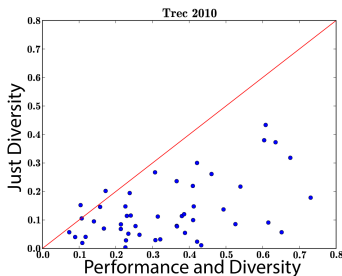
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Per-query conditional mutual information

Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

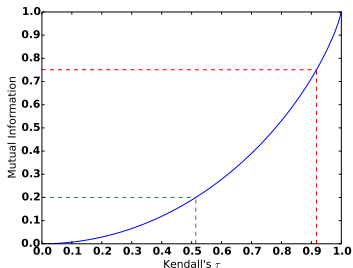
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- Performance and Diversity
- Just Diversity

Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

	TREC 2010	TREC 2011
$\tau_I(\text{ERR-IA} ; \text{D\#-nDCG})$	0.64	0.55
$\tau_I(\text{ERR-IA} ; \text{D\#-nDCG} \mid \text{nDCG})$	0.30	0.17
$\tau_I(\text{ERR-IA} ; \text{D\#-nDCG} \mid \text{ERR})$	0.12	0.14
$\tau_I(\text{ERR-IA} ; \text{D\#-nDCG} \mid \text{nDCG, ERR})$	0.12	0.10

Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

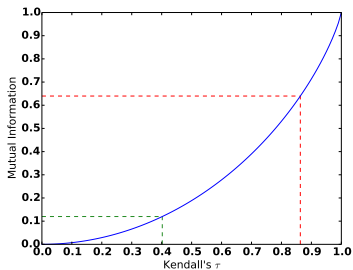
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- $\tau_I(\text{ERR-IA} ; D\#\text{-nDCG})$
- $\tau_I(\text{ERR-IA} ; D\#\text{-nDCG} | \text{ERR}, \text{nDCG})$

Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

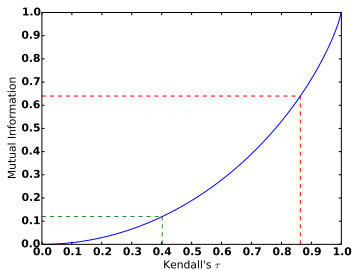
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- $\tau_I(\text{ERR-IA} ; D\#\text{-nDCG})$
- $\tau_I(\text{ERR-IA} ; D\#\text{-nDCG} | \text{ERR}, \text{nDCG})$

Rank-correlation in diversity measures not due to diversity!

Contribution

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Probabilistic Framework for Evaluation and Rank Correlation

- Evaluation and meta-evaluation
- Information-theoretic interpretation
- Many novel applications

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation
- 4 Information Difference**
 - Similarity Classifier
 - Selecting Systems for Metasearch
- 5 Conclusion

Detecting Differences between Rankers

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

How do you know when two systems are different?

- Measure performance delta

Detecting Differences between Rankers

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

How do you know when two systems are different?

- Measure performance delta

What if their performance is the same?

- 1 They are essentially the same

Detecting Differences between Rankers

How do you know when two systems are different?

- Measure performance delta

What if their performance is the same?

- 1 They are essentially the same
- 2 Documents with the same grades appear at the same ranks

Detecting Differences between Rankers

How do you know when two systems are different?

- Measure performance delta

What if their performance is the same?

- 1 They are essentially the same
- 2 Documents with the same grades appear at the same ranks

Want to compare system *behavior* not system *performance*

Information Difference

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

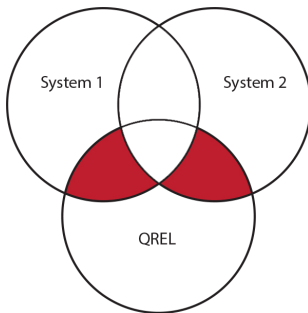
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



$$id(S_1, S_2) = I(S_1; Q \mid S_2) + I(S_2; Q \mid S_1)$$

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

2 Targeted Meta-Evaluation for Understanding Diversity

3 Probabilistic Framework for Evaluation and Rank Correlation

4 Information Difference

- Similarity Classifier
- Selecting Systems for Metasearch

5 Conclusion

“Similarity”

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Research groups submit multiple systems
- Sort systems into bins by performance
- In each bin, mark systems from same group “similar”

ID vs Performance Delta

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

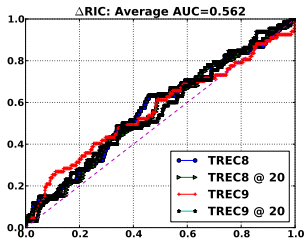
Framework

RIC
Conditional
Rank Correlation

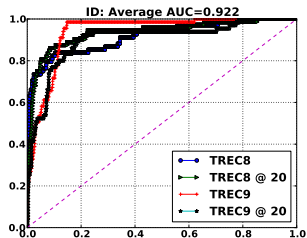
ID

Classifier
Metasearch

Conclusion



Performance Delta



Information Difference

Other Baselines

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Mutual Information: $I(R_{S_1}; R_{S_2})$
 - Order information, no relevance judgments
 - Base case: Kendall's τ
- 2 Jaccard Coefficient: $\frac{S_1 \cap S_2}{S_1 \cup S_2}$
 - No order information, no relevance judgments

Results

Meta-Evaluating Search Engines

Golbus

Motivation

Diversity

Framework

RIC

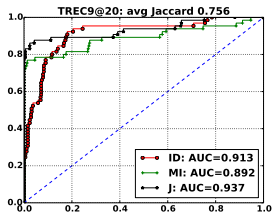
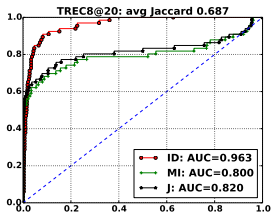
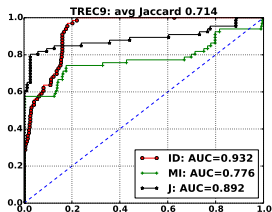
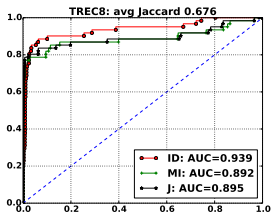
Conditional Rank Correlation

ID

Classifier

Metasearch

Conclusion



Results

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

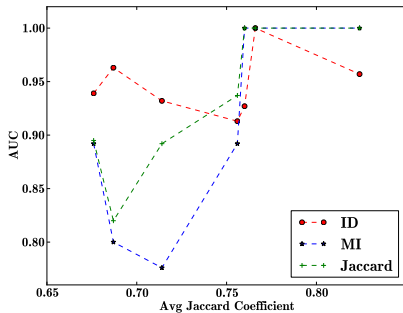
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

2 Targeted Meta-Evaluation for Understanding Diversity

3 Probabilistic Framework for Evaluation and Rank Correlation

4 Information Difference

- Similarity Classifier
- Selecting Systems for Metasearch

5 Conclusion

Metasearch

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Metasearch: combine multiple search engines into a single ranked list that is better than the worst input system.

Which Systems?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

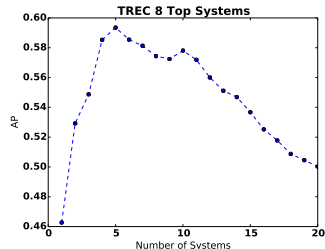
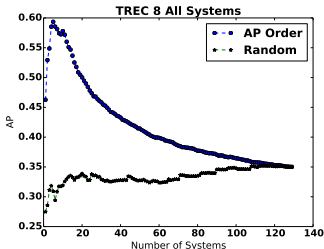
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Impact of system choice

Which Systems?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

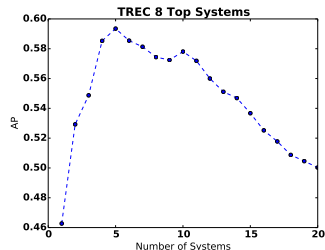
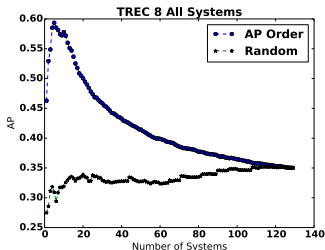
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Impact of system choice

Related work: tail not max

How to choose best without relevance judgments?

Which Systems?

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

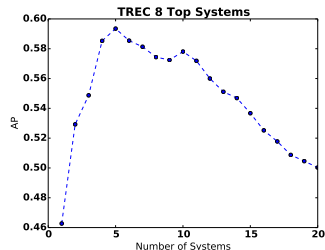
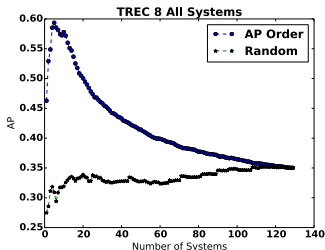
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Impact of system choice

Related work: tail not max

How to choose best without relevance judgments?

Information difference framework

Approach

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Common wisdom: best, most dissimilar systems

- 1 How to define best and most dissimilar?
 - Best: “gold standard” rankers
 - Similarity: **information difference framework**
- 2 How to merge best and most dissimilar?

Facilities Location Analysis

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

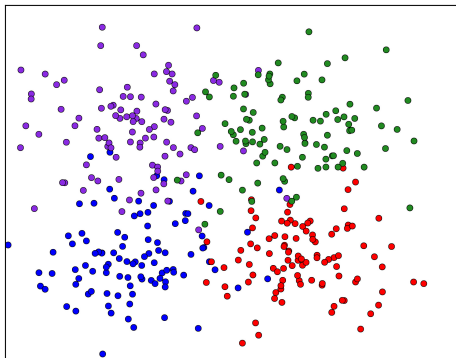
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Synthetic data with 4 clusters

Facilities Location Analysis

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

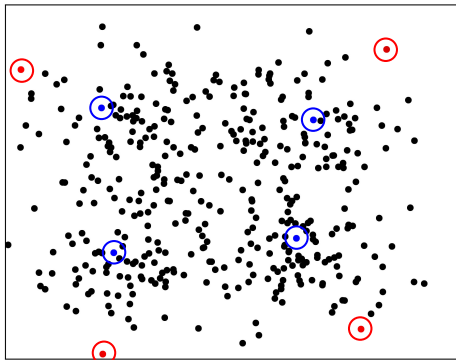
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- **Obnoxious Facilities Dispersion**

- Maximal Marginal Relevance—Most unique*

- **Desirable Facilities Placement**

- *k*-Medoids Clustering—Most representative

*Carbonell & Goldstein. SIGIR '98.

Facilities Location Analysis

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

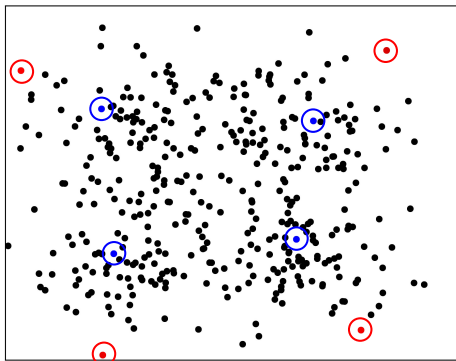
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



- **Obnoxious Facilities Dispersion**

- Maximal Marginal Relevance—Most unique

- **Desirable Facilities Placement**

- *k*-Medoids Clustering—Most representative*

*Zuccon *et al.* ECIR '12

Selecting Systems Without Relevance Judgments

Meta-Evaluating Search Engines

Golbus

Motivation

Diversity

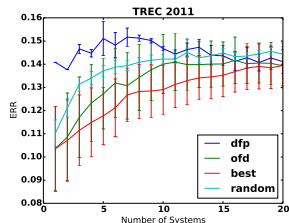
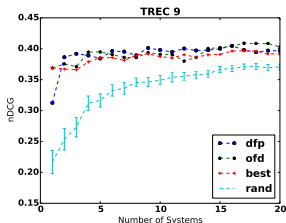
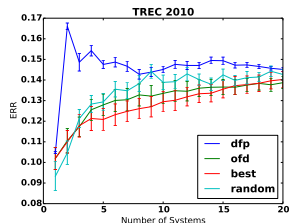
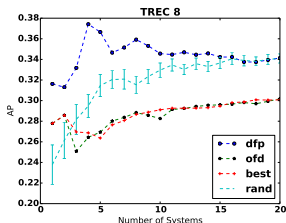
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Best, most *representative* systems

Contribution

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Information Difference

Meta-evaluation tool for comparing systems by *behavior*.

Outline

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- 1 Motivation: Evaluating Information Retrieval
- 2 Targeted Meta-Evaluation for Understanding Diversity
- 3 Probabilistic Framework for Evaluation and Rank Correlation
- 4 Information Difference
- 5 Conclusion**

- 1 Targeted Meta-Evaluation for Understanding Diversity
 - Document Selection Sensitivity
- 2 Probabilistic Framework for Evaluation and Rank Correlation
 - Information-Theoretic Evaluation Measure
 - Conditional Rank Correlation
- 3 Information Difference
 - Similarity Classifier
 - Selecting Systems for Metasearch

ID for measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Information difference: IR specific rank correlation between ranked lists of *documents*
- Can we do this for ranked lists of *systems*, *i.e.* evaluation measures?

ID for measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Information difference: IR specific rank correlation between ranked lists of *documents*
- Can we do this for ranked lists of *systems*, *i.e.* evaluation measures?
- What is the analogue of a QREL?

ID for measures

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

- Information difference: IR specific rank correlation between ranked lists of *documents*
- Can we do this for ranked lists of *systems*, *i.e.* evaluation measures?
- What is the analogue of a QREL?
- Potential application: Comparing crowdsourced QRELs

Mixed Relevance Information

Relevance Judgments

- Experts vs crowdworkers
- Grades vs preferences
 - Direct vs observed
- Estimates from rankers

Mixing Relevance Judgments

Given n QREs, how do we use all of them?

- Currently, use evaluation pipeline: e.g. experts then clicks

Our framework: $I(S; Q_1, \dots, Q_n)$

- Evaluate with respect to all QREs simultaneously
- Each QREL weighted appropriately

Thank You

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Thank you for coming.

Thank You

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

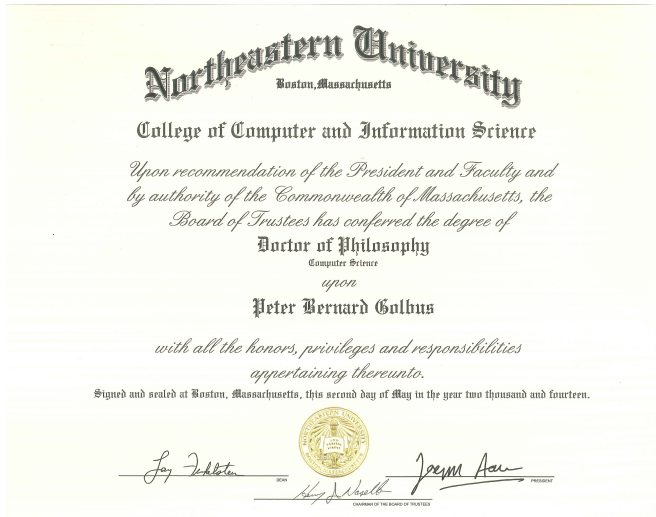
Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion



Thank You

Meta-
Evaluating
Search
Engines

Golbus

Motivation

Diversity

Framework

RIC
Conditional
Rank Correlation

ID

Classifier
Metasearch

Conclusion

Questions / Comments / Concerns