A Mutual Information-based Framework for the Analysis of Information Retrieval Systems

Peter B. Golbus Javed A. Aslam

Northeastern University
College of Computer and Information Science

SIGIR 2013

Outline

Detecting Differences between Rankers

How do you know when two systems are different?

- Measure performance delta

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

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. The documents they retrieved happen to have the same grades

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. The documents they retrieved happen to have the same grades

Can we tell which?

Novelty & Diversity

Subtopics

What is the information need of the query "jaguar"

- animal, car, os x 10.2?

Novelty & Diversity

Subtopics

What is the information need of the query "jaguar"

- animal, car, os x 10.2?

Definitions

- Novelty: documents do not repeat subtopics
- Diversity: list covers many subtopics

Evaluation

1. Performance: how many good documents did you find?
2. Diversity: how well did you order them?

Example

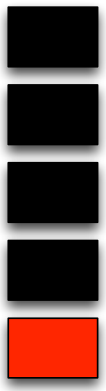
- Subtopic 1
- Subtopic 2
- Subtopic 3
- Not relevant

Golbus, Aslam & Clarke, Increasing evaluation sensitivity to diversity. Inf. Retr. 16(4)

Evaluation

1. Performance: how many good documents did you find?
2. Diversity: how well did you order them?

Bad performance $\Rightarrow$ bad diversity

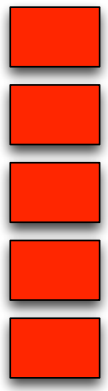


Golbus, Aslam & Clarke, Increasing evaluation sensitivity to diversity. Inf. Retr. 16(4)

Evaluation

1. Performance: how many good documents did you find?
2. Diversity: how well did you order them?

Good performance $\not\Rightarrow$ good diversity

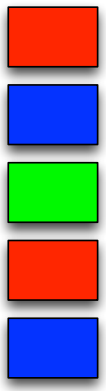


Golbus, Aslam & Clarke, Increasing evaluation sensitivity to diversity. Inf. Retr. 16(4)

Evaluation

1. Performance: how many good documents did you find?
2. Diversity: how well did you order them?

Good diversity $\Rightarrow$ good performance



Golbus, Aslam & Clarke, Increasing evaluation sensitivity to diversity. Inf. Retr. 16(4)

Evaluation

1. Performance: how many good documents did you find?
2. Diversity: how well did you order them?

Evaluating Diversity

- Add diversity component to existing performance measure
- Are we really measuring diversity?

Golbus, Aslam & Clarke, Increasing evaluation sensitivity to diversity. Inf. Retr. 16(4)

Metaevaluation

Hypothesis

If we are truly measuring diversity, then the diversity components of various measures will be correlated.

How do we measure this?

Outline

Observation 1

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

Observation 1

If we reframe gain and discount as random variables, then $eval(S) = E[G \cdot D]$

Observation 1

If we reframe QREL and ranked list as random variables, then $eval(S) = E[Q \cdot R]$

Observation 1

We should let $eval(S) = I(Q; R)$

Observation 2

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

Observation 2

- Search engine: *total* ordering on *some* documents
- QREL: *partial* ordering on *all* documents
- Evaluation: comparison of orderings

How do you compare different types of orderings on partially overlapping objects?

Probabilistic Framework

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

Probabilistic Framework

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

Probabilistic Framework

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

Probabilistic Framework

1. Sample space
   - $\Omega =$ all $2 \cdot \binom{n}{2}$ document pairs
2. Distribution over sample space
   - $P = U$
3. Random variables
   - For a ranked list R, $X_R : \Omega \rightarrow \{-1, +1\}$

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

   - $E[X_R] = 0$

Probabilistic Framework

1. Sample space
   - $\Omega =$ all $2 \cdot \binom{n}{2}$ document pairs
2. Distribution over sample space
   - $P = U$
3. Random variables
   - For a ranked list R, $X_R : \Omega \rightarrow \{-1, +1\}$

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

   - $E[X_R] = 0$

For two lists R and S

- $E[X_R \cdot X_S] = \frac{2C-2D}{2(C+D)}$

Probabilistic Framework

1. Sample space
   - $\Omega =$ all $2 \cdot \binom{n}{2}$ document pairs
2. Distribution over sample space
   - $P = U$
3. Random variables
   - For a ranked list R, $X_R : \Omega \rightarrow \{-1, +1\}$

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

   - $E[X_R] = 0$

For two lists R and S

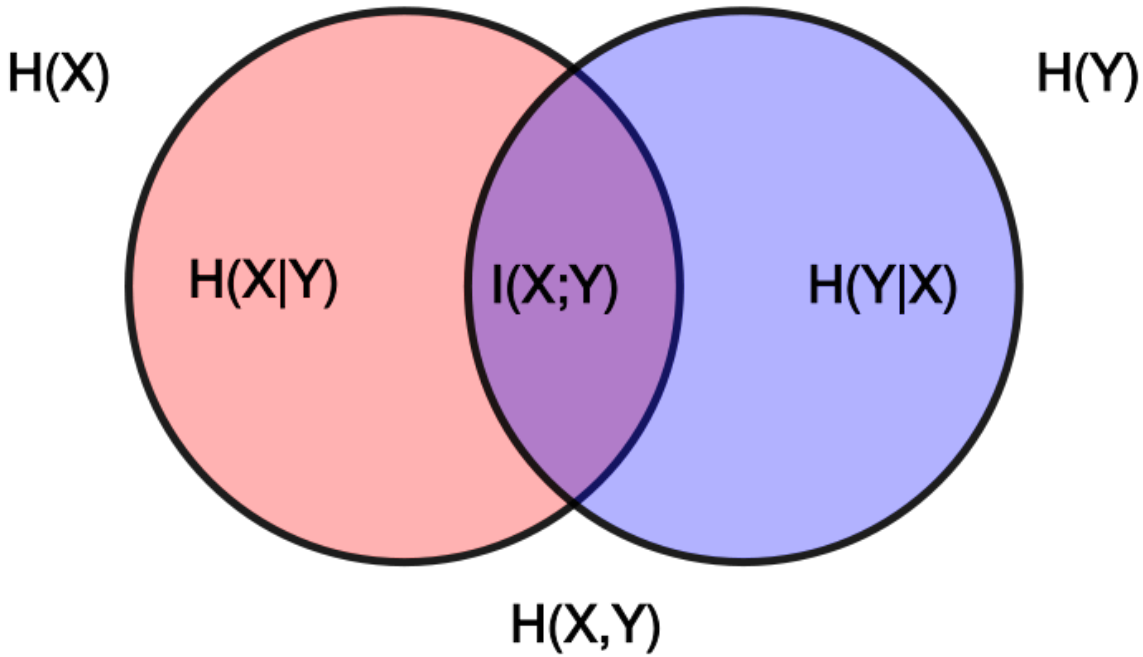
- $E[X_R \cdot X_S] = \frac{2C-2D}{2(C+D)} = \tau(R, S)$

Power of Framework

Power of Framework

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

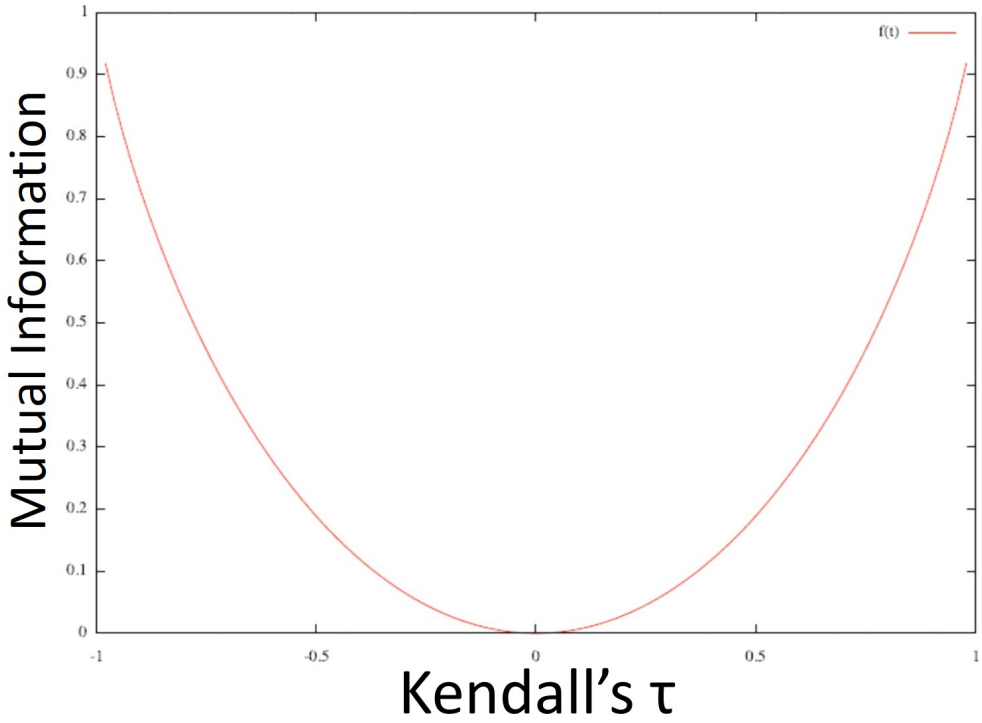
Information Theory



$I(S_1, S_2; Q)$, $I(S_1, \ldots S_n; Q)$, $(S; Q_1, \ldots, Q_n)$, $\ldots$

Outline

Application



Conditional Rank Correlation

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

Metaevaluation

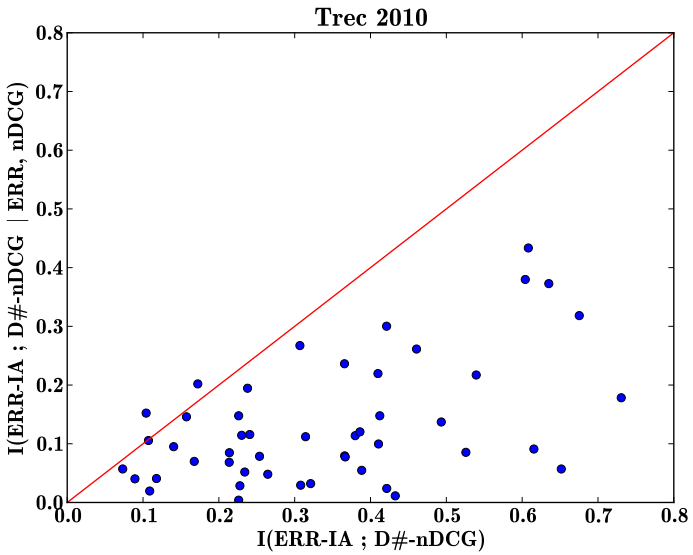
Hypothesis

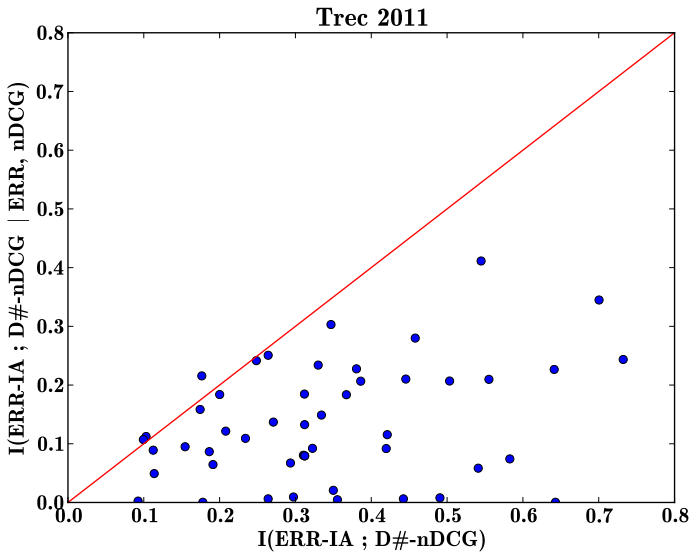
If we are truly measuring diversity, then the diversity components of various measures will be correlated.

Experiment

If $\tau_I(D\#\text{-meas}, \text{meas-}IA)$ is large, then $\tau_I(D\#\text{-meas}, \text{meas-}IA \mid meas)$ should also be large.

Results





per query conditional mutual information

Results

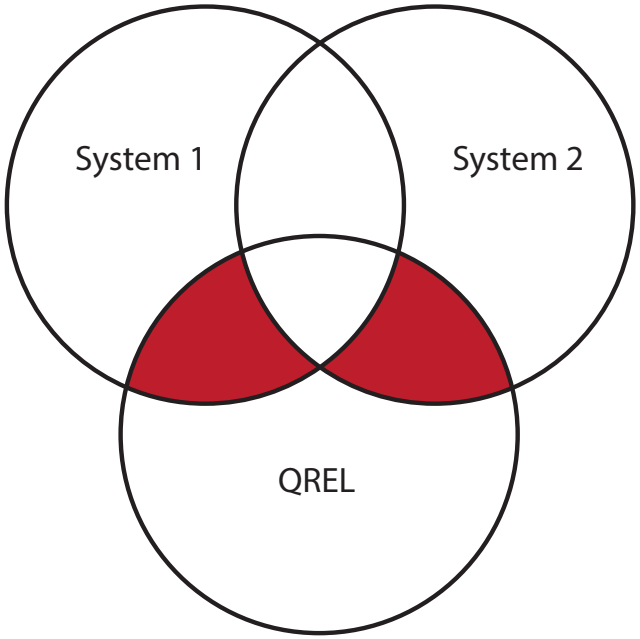
	TREC 2010	TREC 2011
τ_I(ERR-IA ; D#-nDCG)	0.6390	0.5545
τ_I(ERR-IA ; D#-nDCG | nDCG)	0.3026	0.1728
τ_I(ERR-IA ; D#-nDCG | ERR)	0.1222	0.1442
τ_I(ERR-IA ; D#-nDCG | nDCG, ERR)	0.1239	0.1003

Outline

Information Difference

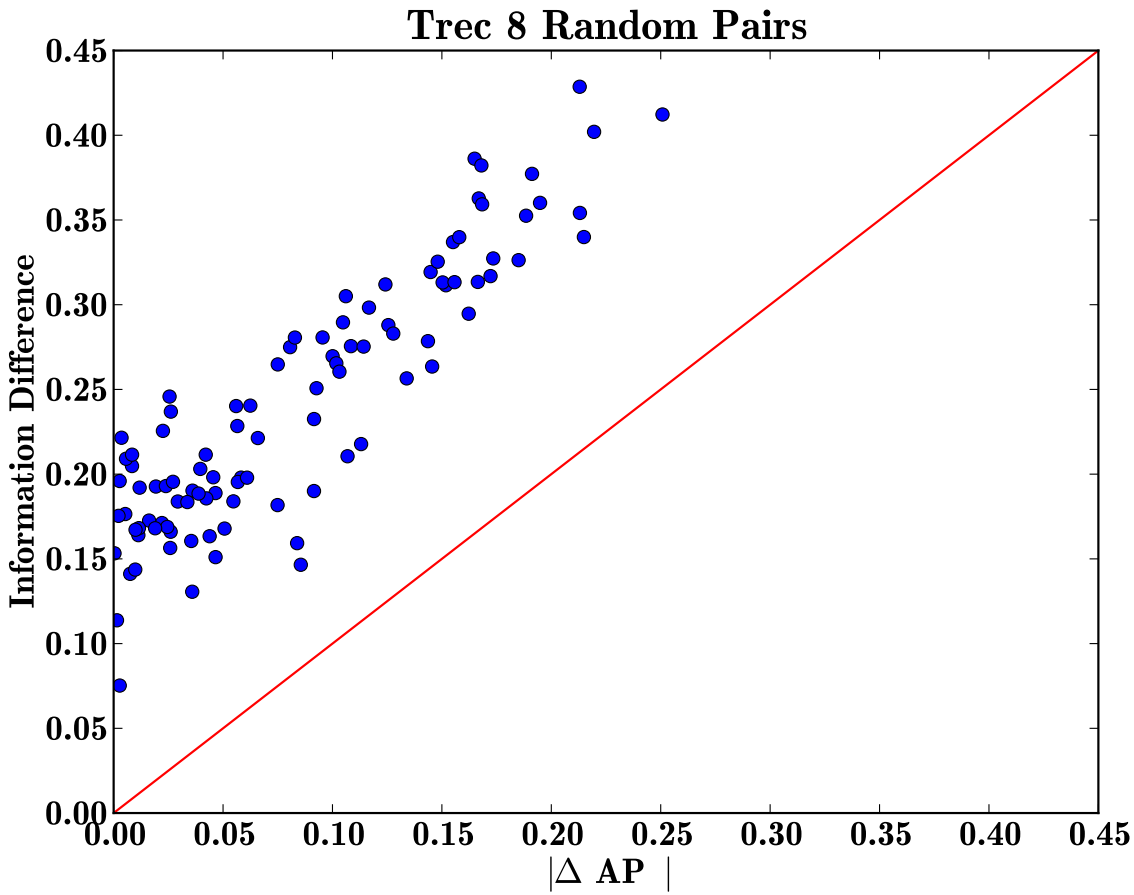
How can we detect differences between systems with the same performance?

Information Difference



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

“Sniff Test”



Experiment

Hypothesis

Similar systems will have small information difference.

Experiment

Research groups submit multiple systems. Are system pairs with (1) similar performance and (2) small information difference from the same group?

- Sort systems into bins by AP
- For each bin:
  - Compute information difference for each system pair
- Sort by *id*

Results

Rank	System 1	System 2	*id*	$\lvert\Delta$ AP$\rvert$
1	UB99T	UB99SW	0.010	0.005
2	unc8al32	unc8al42	0.012	0.002
3	fub99tt	fub99tf	0.017	0.000
4	nttd8al	nttd8alx	0.023	0.002
5	ibmg99a	ibmg99b	0.027	0.012
		⋮		
28	isa25t	cirtrc82	0.084	0.004
29	CL99SD	CL99SDopt2	0.086	0.000
30	ok8amxc	ok8alx	0.086	0.006
31	tno8d4	MITSLStd	0.088	0.016
32	uwmt8a2	uwmt8a1	0.089	0.002

Outline

Future Work

In the paper:

- Evaluation Measure
- Upper Bound on Metasearch

Future Work

- Rank Correlation & Preference Aggregation
- Diversity Evaluation
- Web-Scale Evaluation with Mixed Relevance Information

Thank You

Questions / Comments / Concerns?