Detecting Duplicate Web Documents using Clickthrough Data

Filip Radlinski,
Paul N. Bennett, Emine Yilmaz

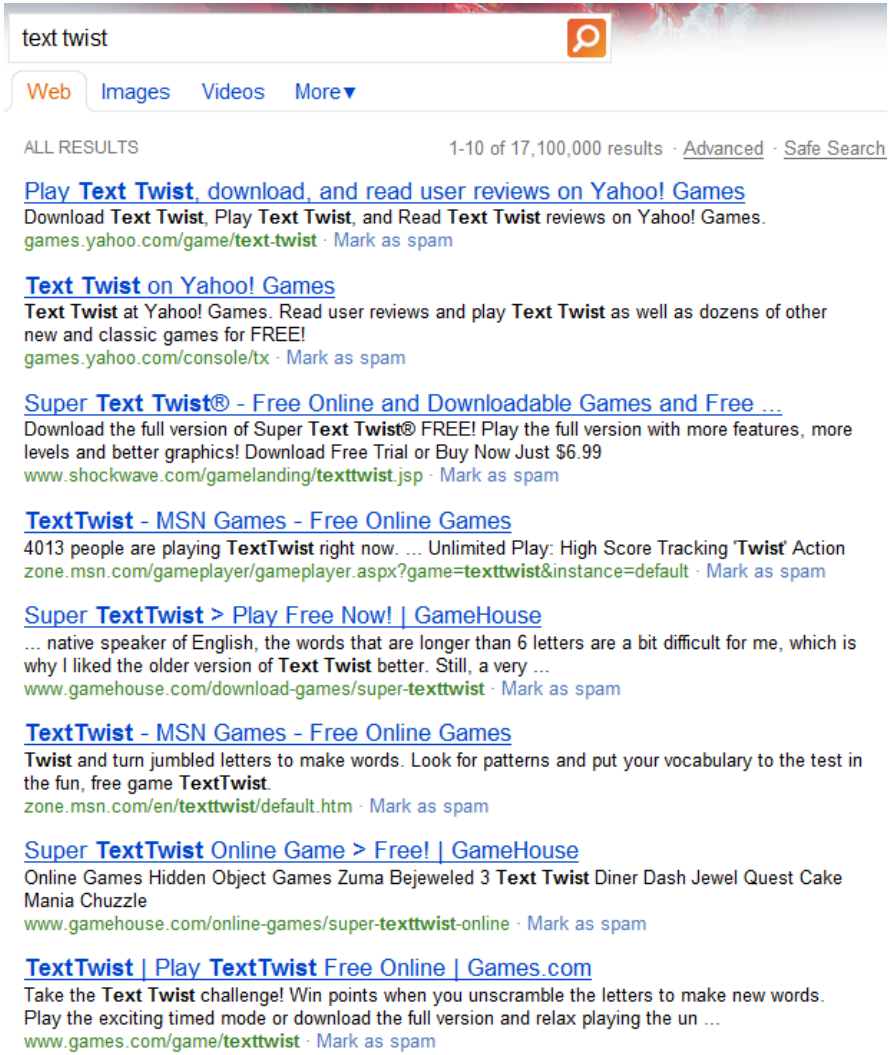
Microsoft

Motivation

- Near-Duplicates and duplicates are common.
- In many cases, showing them isn't helpful.
- Usually identified based on page *content*.

➢ not sensitive to query.

- We show that clicks give a more context-sensitive duplication signal.



Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Previous approaches

- Detecting Duplication using Content
  - Fingerprints for exact duplicates:

  Compute a fingerprint for each document. If fingerprints match, check if the documents are the same.

  - Shingling to find near duplicates:

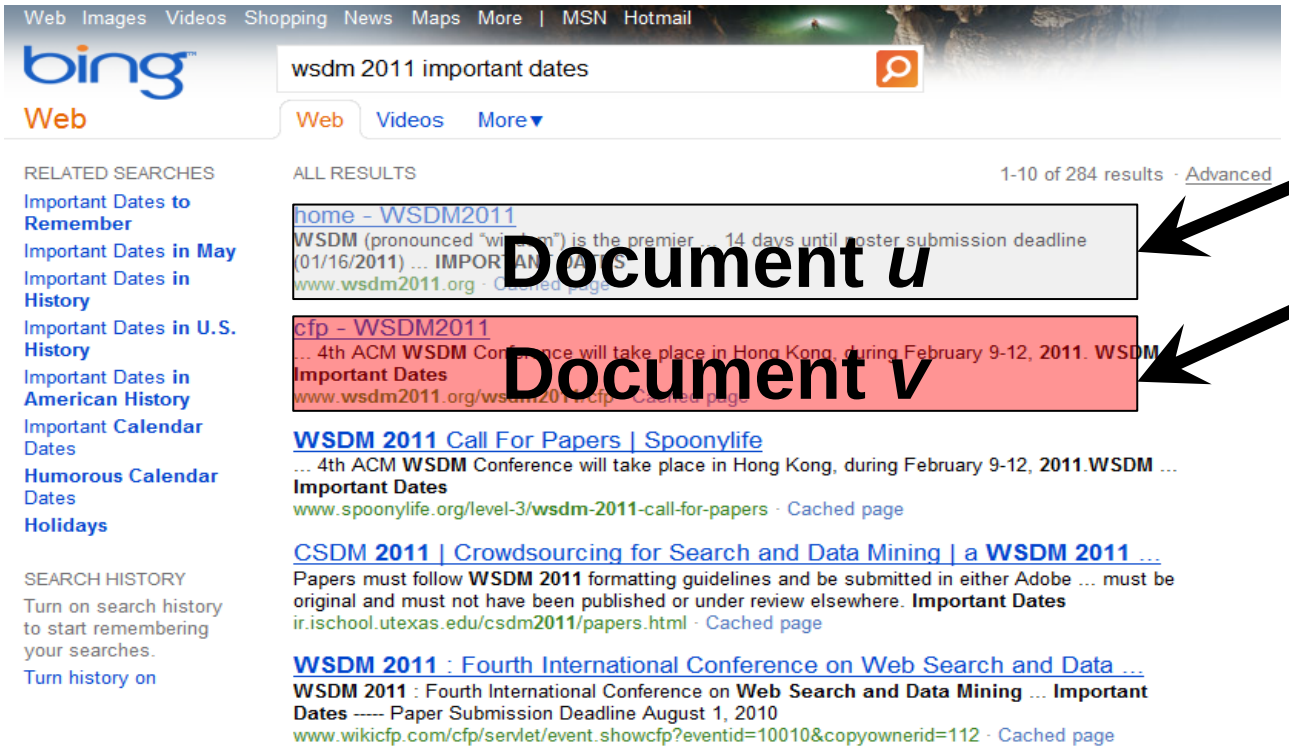
  For each ngram, compute a fingerprint. Measure the similarity between a summary of fingerprints.

- Reducing redundancy in search results
  - Given relevance and similarity, rank results by relevance minus redundancy

- These methods don't depend on the query.

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score



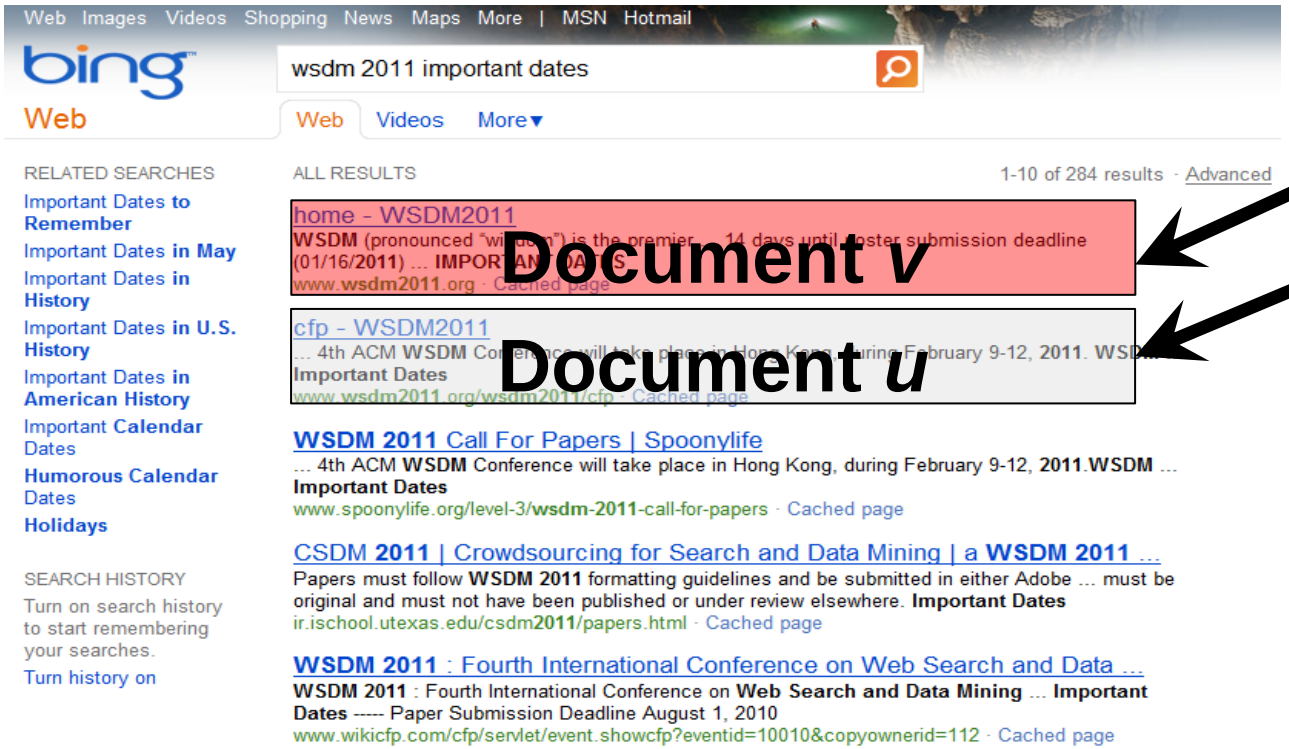
Click rate $c^{\hat{u}v}$

Click rate $c^{u\hat{v}}$

$$bias_{uv} = \frac{c^{\hat{u}v}}{c^{\hat{u}v} + c^{u\hat{v}} + c^{\hat{u}\hat{v}}}$$

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score



$$bias_{vu} = \frac{c^{\hat{v}u}}{c^{\hat{v}u} + c^{v\hat{u}} + c^{\hat{v}\hat{u}}}$$

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score

- The redundancy score is the minimum of these two ratios:

$$r(u,v) = \min\left(\frac{c^{\hat{u}v}}{c^{\hat{u}v} + c^{u\hat{v}} + c^{\hat{u}\hat{v}}}, \frac{c^{\hat{v}u}}{c^{\hat{v}u} + c^{v\hat{u}} + c^{\hat{v}\hat{u}}}\right)$$

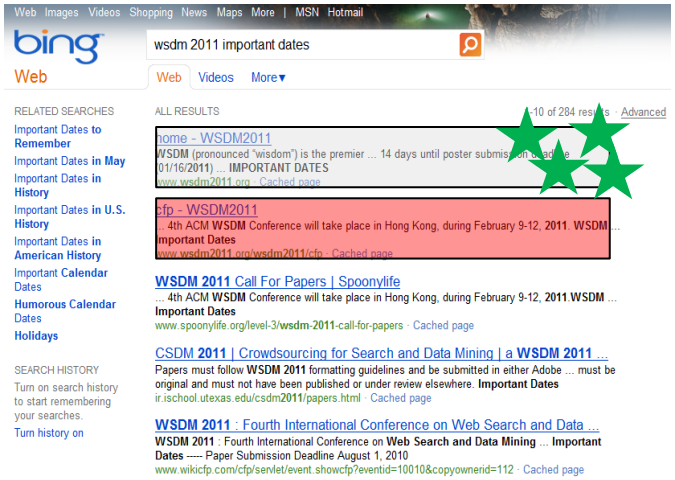
- In words:

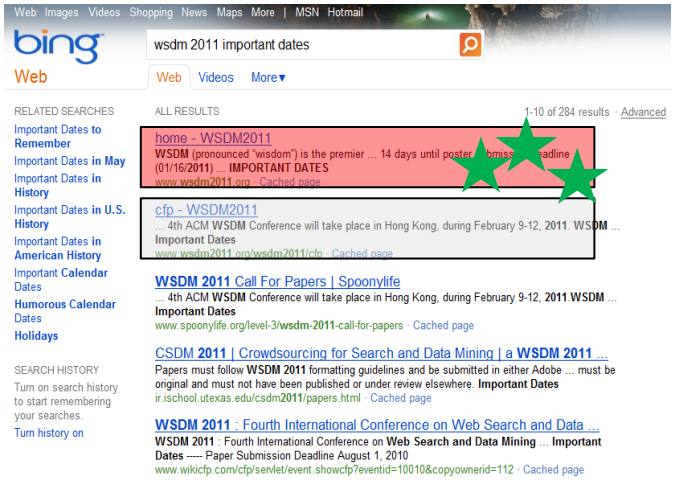
  Across **both** presentation orders, what is the **minimum** rate at which just the **top** result is clicked?

- High presentation bias $\longleftrightarrow$ High redundancy score

Redundancy Score

- If in both orders, the top result always gets clicked:





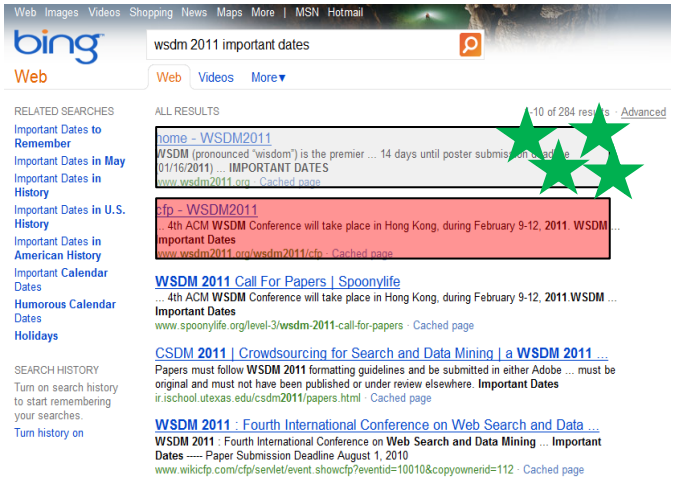
➢ The results are probably duplicate

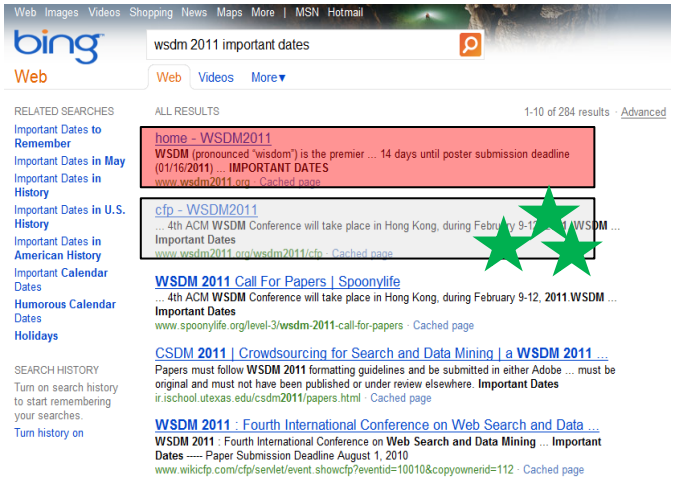
  ➢ Or users always just click on the top result: we'll check

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score

- If one of the results is always clicked on, even when its lower:



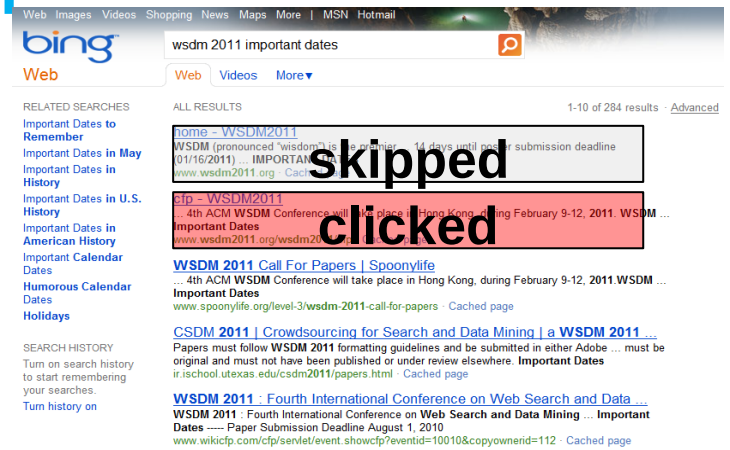


➢ That result is preferred, these are not redundant

Most real document pairs are in between

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Relationship to Click-Skip and FairPairs



- The Click-Skip approach:

  A clicked document is more relevant than skipped ones above

- FairPairs:

  If you present documents in both orders, the one with higher bottom click rate is more relevant.

  ➔ A ***bottom click*** is a ***relevance*** signal

- This work:

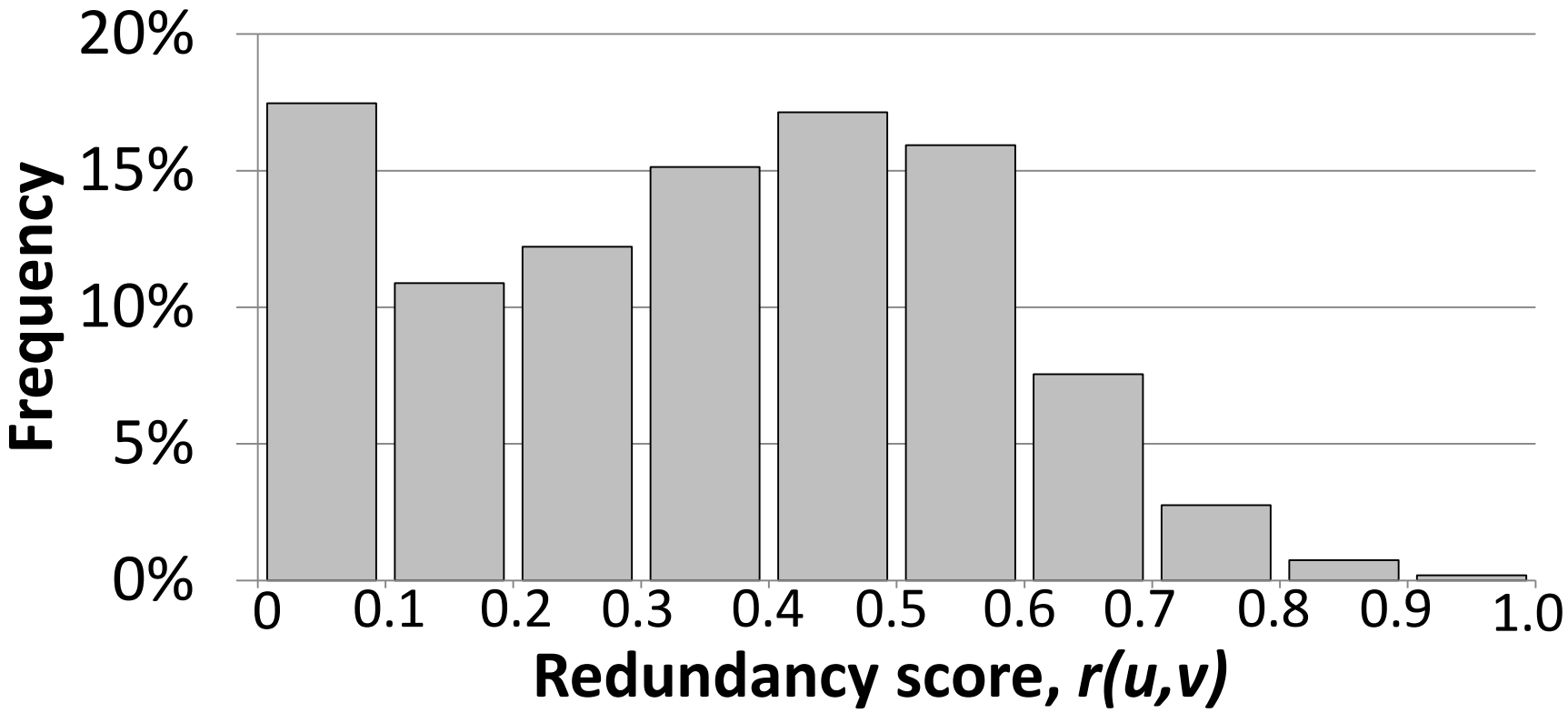
  Present documents in both orders. If the top one is always clicked, the documents are duplicate.

  ➔ A ***top click*** is a ***duplication*** signal

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Score Distribution

- Does real document pairs exist with a variety of redundancy scores?



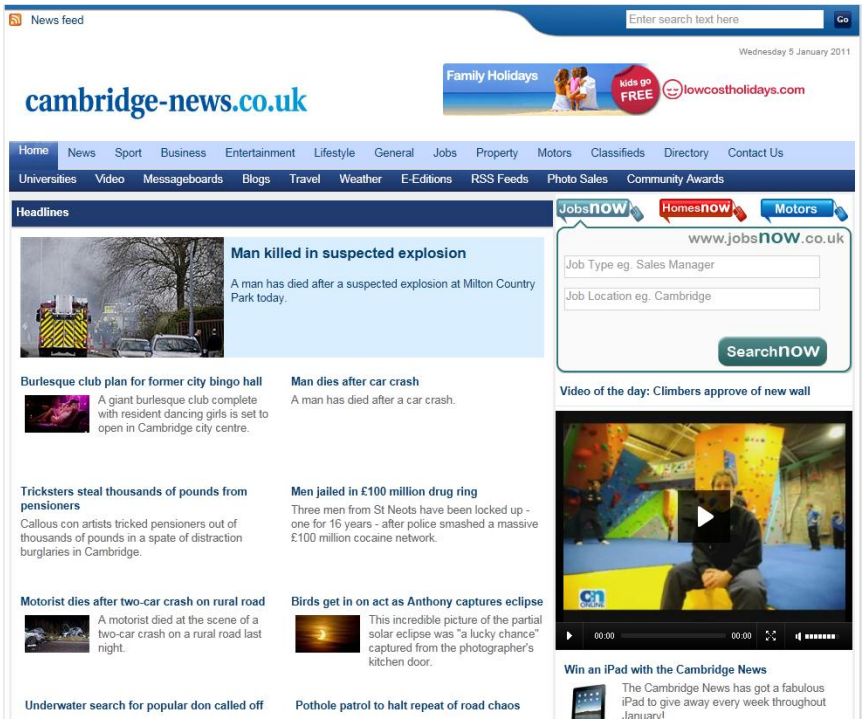
Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

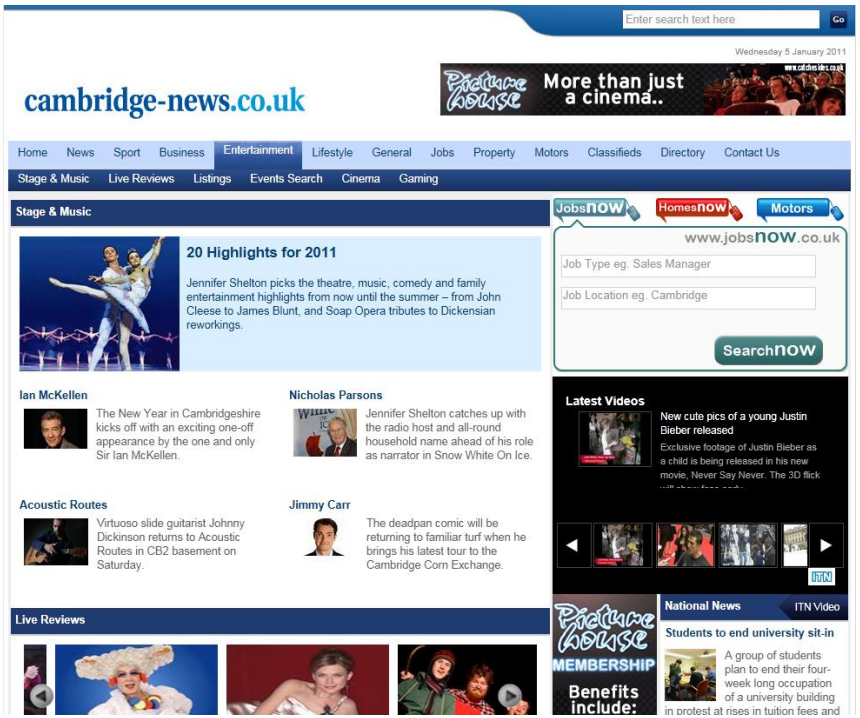
Classes of Duplication

Inspecting pairs of documents with high redundancy score, three types of duplicates jump out:

- **Exact duplicates**: Both pages appear identical, perhaps with the exception of ads.
- **Content duplicates**: Both pages provide the same / very related information (for this query), but from different sources
- **Navigational duplicates**: Getting from one page to the other is very easy

Example: Navigational Duplicates

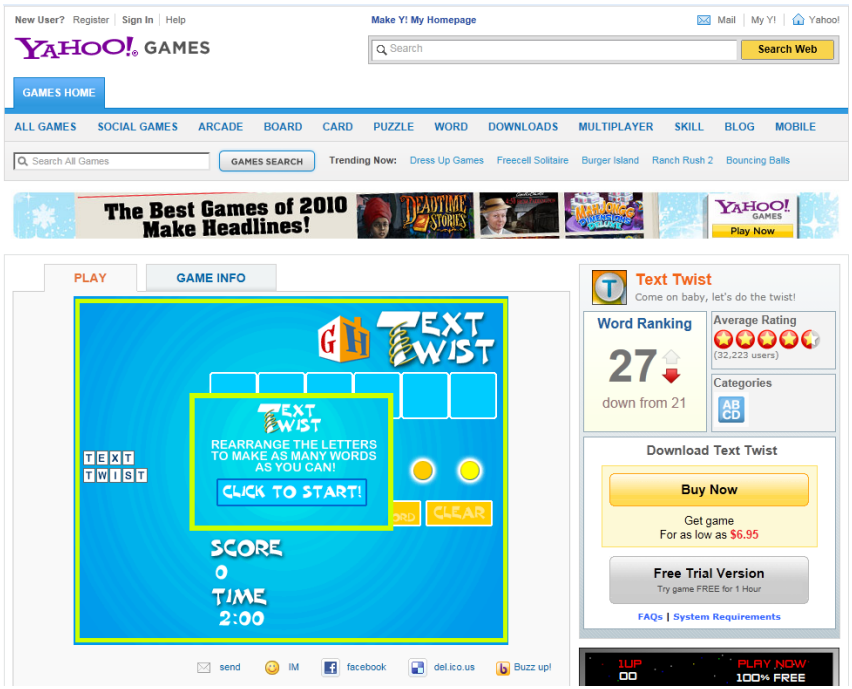


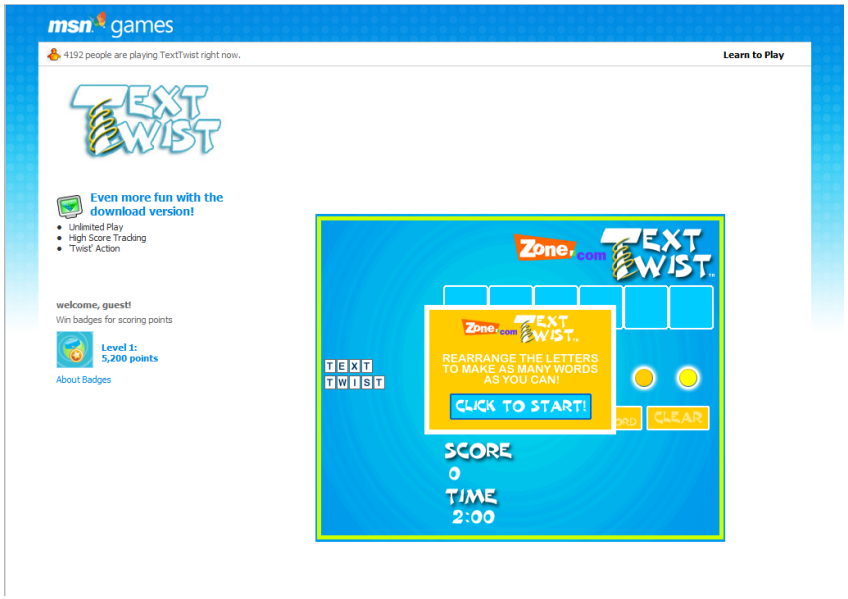


- Other common examples:
  - Bank homepage vs. online banking page
  - Related Amazon products or eBay auctions

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Example: Content Duplicates





- Other common examples:
  - Competing song lyrics websites
  - Different recipe websites
  - Competing sofa manufacturers

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Evaluation Approach

- Test if redundancy score tells us about duplication
  1. Sample tuples with variety of redundancy scores
  2. Judge the (query, url, url) triplets for duplication
  3. Measure agreement between score and judgments
  4. Train duplicate classifiers

- For each triplet, asked three questions:
  1. Which page is most relevant to the query?
  2. How similar is the utility of these pages for the query?
  3. Is it “easy” to navigate from either page to the other?

Judging Duplication

More Relevant?	Utility?	Navigation?	Freq.	
Both equally	Identical	-	6%	Exact duplicates
Both equally	*any*	Yes within	5%	Navigational duplicates
Left or Right	*any*	Yes within	13%	
Both equally	Very similar	No	16%	Content duplicates
Left or Right	Very similar	No	8%	
Both equally	Related	No	4%	"Weak" content duplicates
Left or Right	Related	No	15%	
Different intents	-	No	12%	Not Duplicate
Left or Right	Different	No	12%	
Left/Right/Both	*any*	Yes across	2%	
Neither Relevant	*any*	*any*	4%	
other	*other*	*other*	3%	

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Judging Duplication

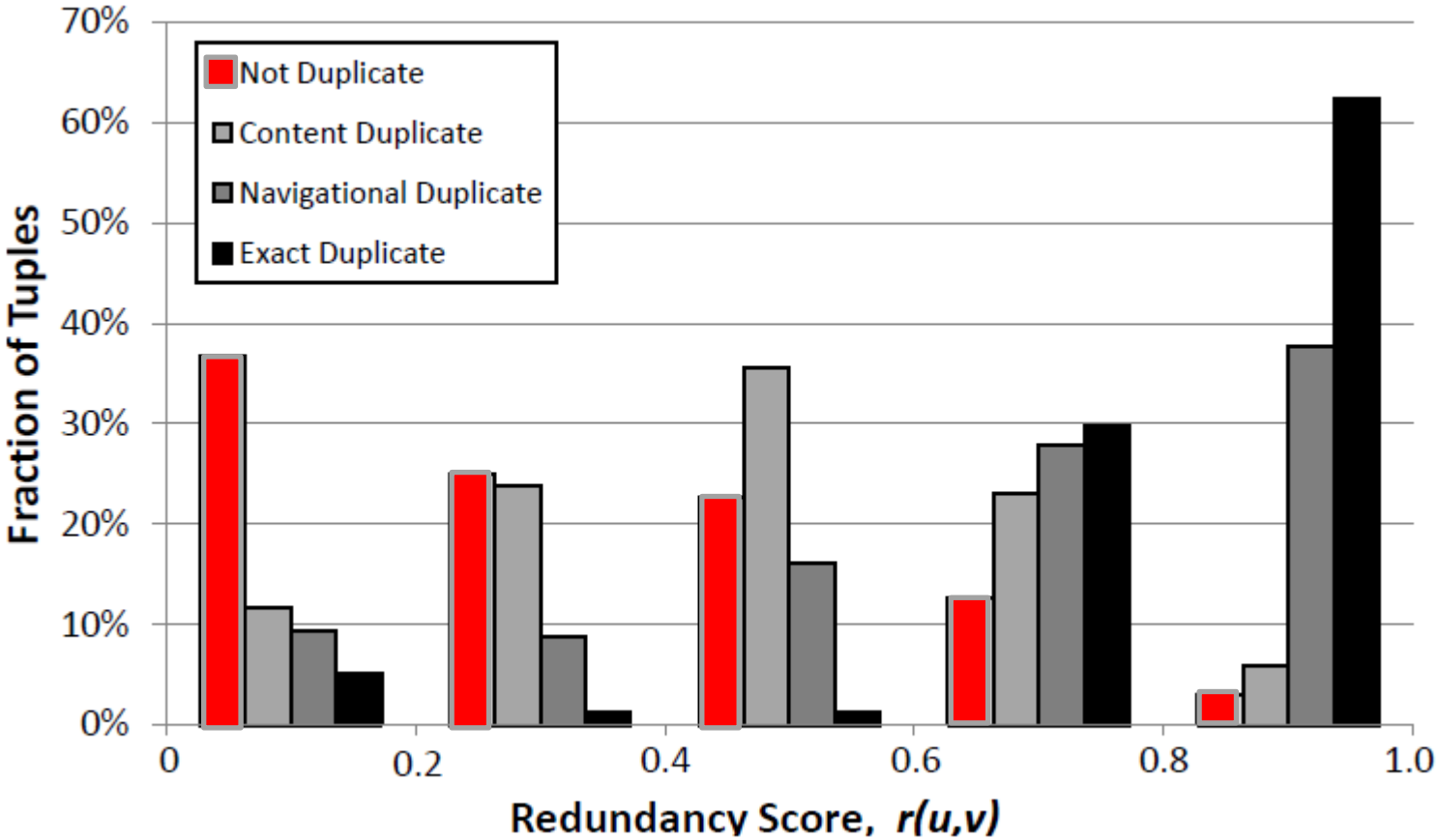
- Inter-judge agreement on a small set tells us its tricky to make these judgments

	Judgment 1				
Judgment 2	Ex.	Nav.	Cont.	C_w	Not Dup.
Exact (E)	0	0	0	0	0
Navigational (N)		19	1	3	2
Content (C)			31	27	12
Weak Cont. (C_w)				15	32
Not Duplicate					78

- The hard one is content vs. weak content vs. not duplicate.
  - The judgments often differ by one level

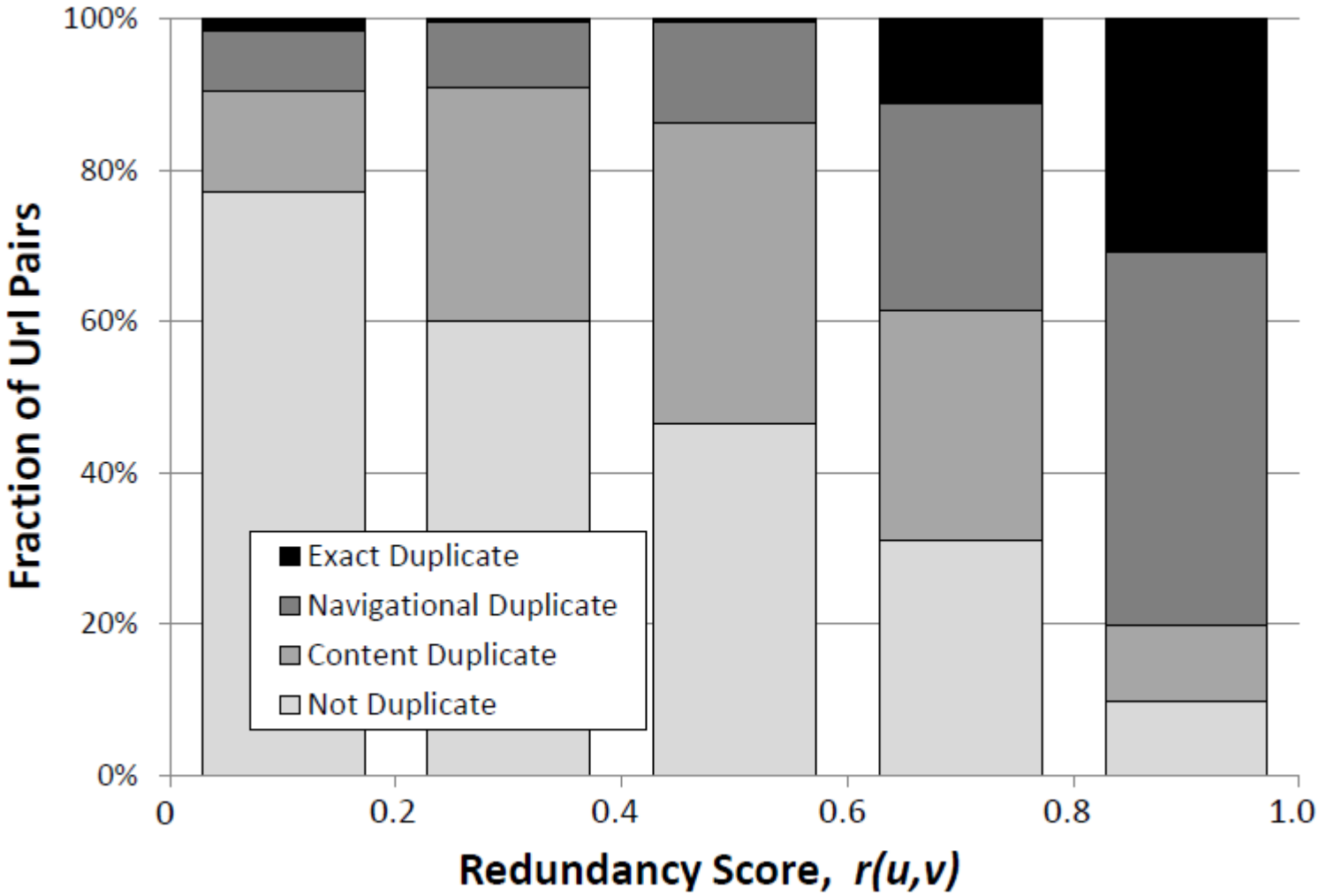
Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Judgments vs Redundancy Score



Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

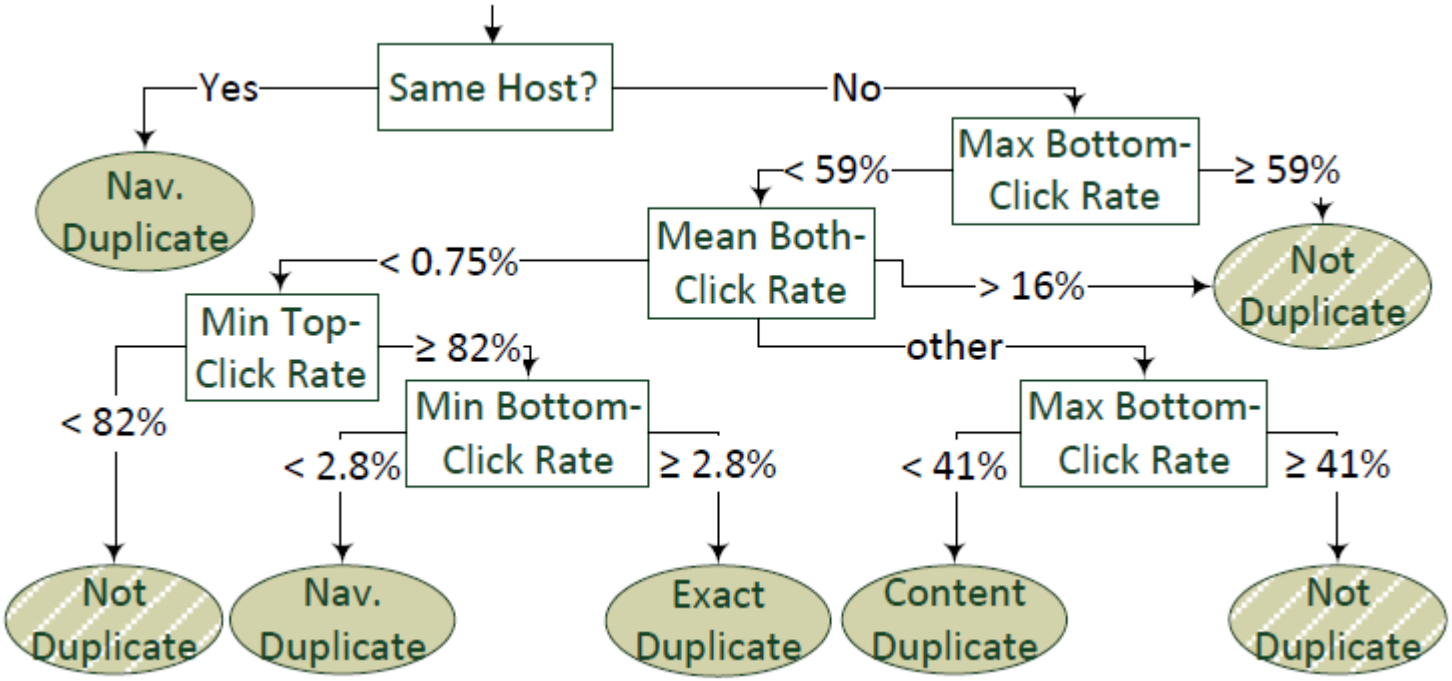
Judgments vs Redundancy Score



Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

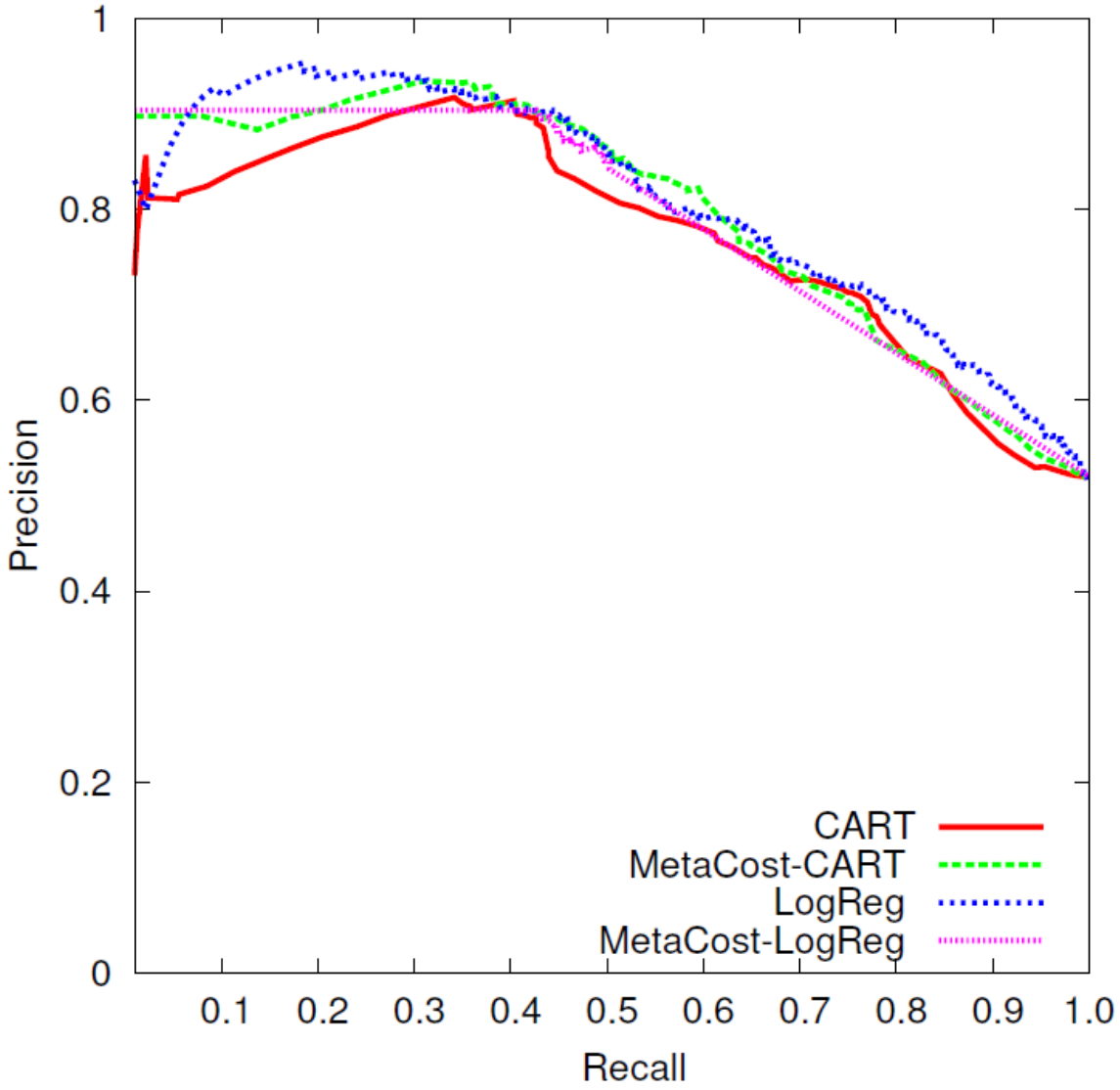
Learning Duplicate Detection

- Training on all our data, different click scores distinguish between the different classes:



Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Learning Duplicate Detection



Bennett & Yilmaz, WSDM 2011

Acting on Duplication

- Assuming that we can detect the classes of duplicates, what should we do?
  - Exact duplicates: Remove them.
  - Navigational duplicates:
    - Probably pick just one, but the right one!
  - Content duplicates:
    - Maybe tweak the UI to show them as alternatives?
- Beyond modifying search result rankings:
  - Better relevance from clicks
  - Clean up training/evaluation data

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Conclusions & Open Questions

- We proposed a taxonomy of duplication.
- Clicks can be used to distinguish among the classes.
- Presentation bias has a limited effect on non-duplicates.

- Sometimes (near-)duplicates are useful. How can we measure how useful they are?
- How to obtain more reliable evaluation data?
- Investigate how other duplication signals (e.g. content) help classification.