

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.

The screenshot shows a Bing search results page for the query "text twist". The search bar at the top contains the text "text twist" and a magnifying glass icon. Below the search bar are tabs for "Web", "Images", "Videos", and "More". The results section is titled "ALL RESULTS" and shows "1-10 of 17,100,000 results". The first result is "Play Text Twist, download, and read user reviews on Yahoo! Games" with a link to "games.yahoo.com/game/text-twist". The second result is "Text Twist on Yahoo! Games" with a link to "games.yahoo.com/console/tx". The third result is "Super Text Twist® - Free Online and Downloadable Games and Free ..." with a link to "www.shockwave.com/gamelanding/texttwist.jsp". The fourth result is "TextTwist - MSN Games - Free Online Games" with a link to "zone.msn.com/gameplayer/gameplayer.aspx?game=texttwist&instance=default". The fifth result is "Super TextTwist > Play Free Now! | GameHouse" with a link to "www.gamehouse.com/download-games/super-texttwist". The sixth result is "TextTwist - MSN Games - Free Online Games" with a link to "zone.msn.com/en/texttwist/default.htm". The seventh result is "Super TextTwist Online Game > Free! | GameHouse" with a link to "www.gamehouse.com/online-games/super-texttwist-online". The eighth result is "TextTwist | Play TextTwist Free Online | Games.com" with a link to "www.games.com/game/texttwist".

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.

Redundancy Score

Web Images Videos Shopping News Maps More | MSN Hotmail

bing™

Web Videos More▼

RELATED SEARCHES

- Important Dates to Remember
- Important Dates in May
- Important Dates in History
- Important Dates in U.S. History
- Important Dates in American History
- Important Calendar Dates
- Humorous Calendar Dates
- Holidays

SEARCH HISTORY

Turn on search history to start remembering your searches.

Turn history on

ALL RESULTS

1-10 of 284 results · [Advanced](#)

Document u

Document v

[WSDM 2011 Call For Papers | Spoonylife](#)
... 4th ACM WSDM Conference will take place in Hong Kong, during February 9-12, 2011. WSDM ...
Important Dates
[www.spoonylife.org/level-3/wdsdm-2011-call-for-papers](#) · [Cached page](#)

[CSDM 2011 | Crowdsourcing for Search and Data Mining | a WSDM 2011 ...](#)
Papers must follow WSDM 2011 formatting guidelines and be submitted in either Adobe ... must be original and must not have been published or under review elsewhere. **Important Dates**
[ir.ischool.utexas.edu/csdm2011/papers.html](#) · [Cached page](#)

[WSDM 2011 : Fourth International Conference on Web Search and Data ...](#)
WSDM 2011 : Fourth International Conference on Web Search and Data Mining ... **Important Dates** ----- Paper Submission Deadline August 1, 2010
[www.wikicfp.com/cfp/servlet/event.showcfp?eventid=10010©ownerid=112](#) · [Cached page](#)

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

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

Redundancy Score

Web Images Videos Shopping News Maps More | MSN Hotmail

bing™

Web Videos More▼

RELATED SEARCHES

- Important Dates to Remember
- Important Dates in May
- Important Dates in History
- Important Dates in U.S. History
- Important Dates in American History
- Important Calendar Dates
- Humorous Calendar Dates
- Holidays

SEARCH HISTORY

Turn on search history to start remembering your searches.

Turn history on

ALL RESULTS 1-10 of 284 results · [Advanced](#)

home - WSDM2011
WSDM (pronounced "wisdom") is the premier ... 14 days until poster submission deadline (01/16/2011) ... **IMPORTANT DATES**
www.wsdm2011.org · [Cached page](#)

cfp - WSDM2011
... 4th ACM WSDM Conference will take place in Hong Kong during February 9-12, 2011. WSDM ... **Important Dates**
www.wsdm2011.org/wsdm2011/cfp · [Cached page](#)

WSDM 2011 Call For Papers | Spoonylife
... 4th ACM WSDM Conference will take place in Hong Kong, during February 9-12, 2011. WSDM ... **Important Dates**
www.spoonylife.org/level-3/wsdm-2011-call-for-papers · [Cached page](#)

CSDM 2011 | Crowdsourcing for Search and Data Mining | a WSDM 2011 ...
Papers must follow WSDM 2011 formatting guidelines and be submitted in either Adobe ... must be original and must not have been published or under review elsewhere. **Important Dates**
ir.ischool.utexas.edu/csdm2011/papers.html · [Cached page](#)

WSDM 2011 : Fourth International Conference on Web Search and Data ...
WSDM 2011 : Fourth International Conference on Web Search and Data Mining ... **Important Dates** ----- Paper Submission Deadline August 1, 2010
www.wikicfp.com/cfp/servlet/event.showcfp?eventid=10010©ownerid=112 · [Cached page](#)

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

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

Redundancy Score

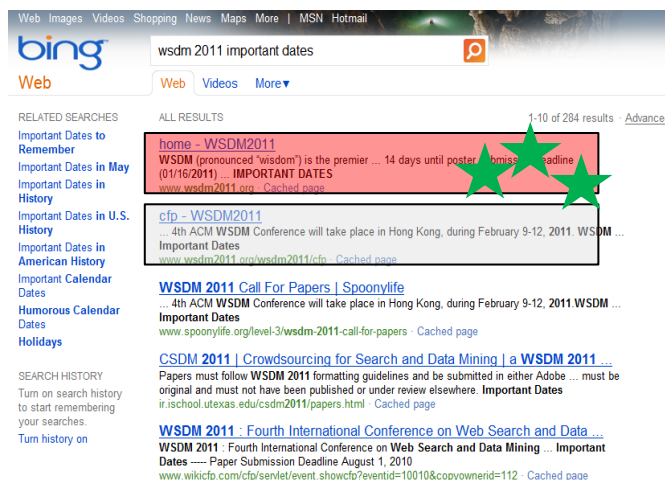
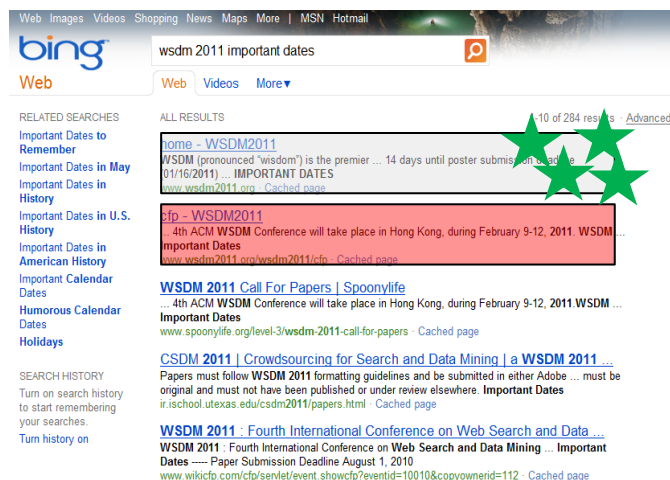
- The redundancy score is the minimum of these two ratios:

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

- 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

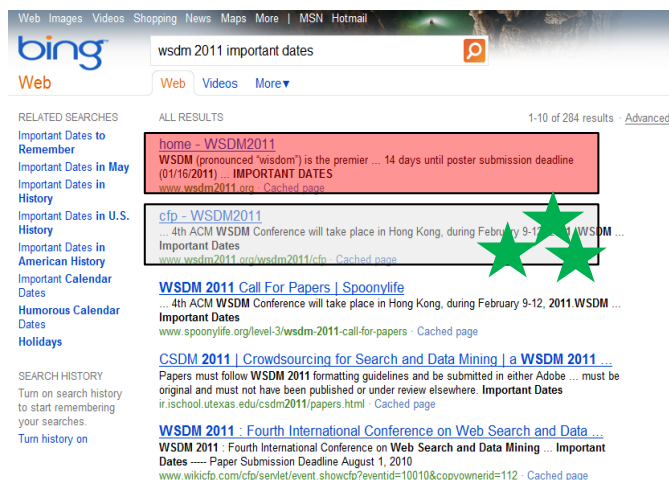
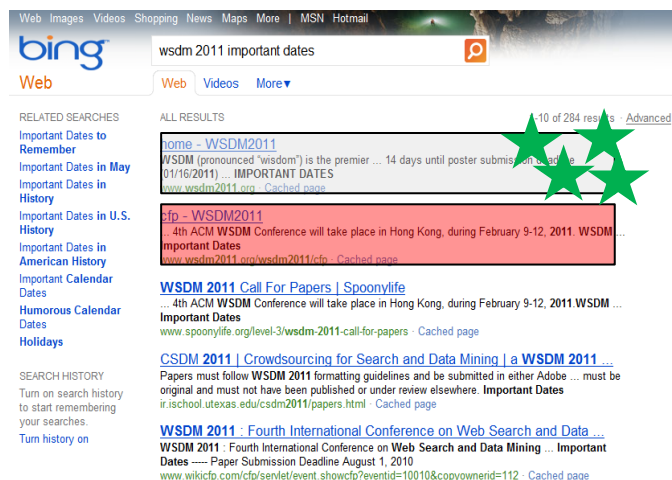
- If in both orders, the top result always gets clicked:



- The results are probably duplicate
- Or users always just click on the top result: we'll check

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

Relationship to Click-Skip and FairPairs

- The Click-Skip approach:

A clicked document is more relevant than skipped ones above

- FairPairs:

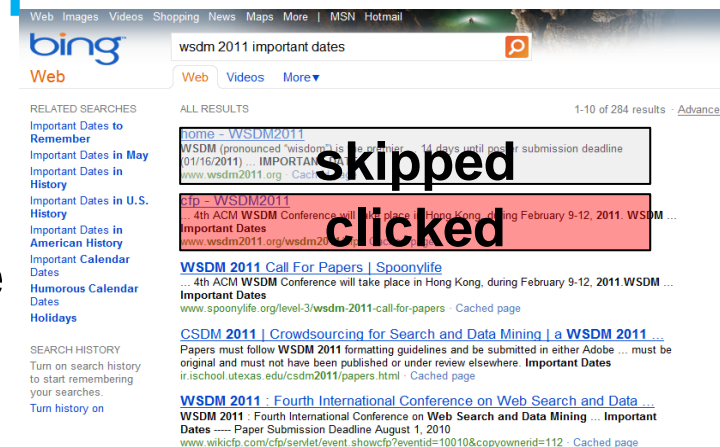
If you present documents in both orders, the one with higher bottom click rate is more relevant.

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

- This work:

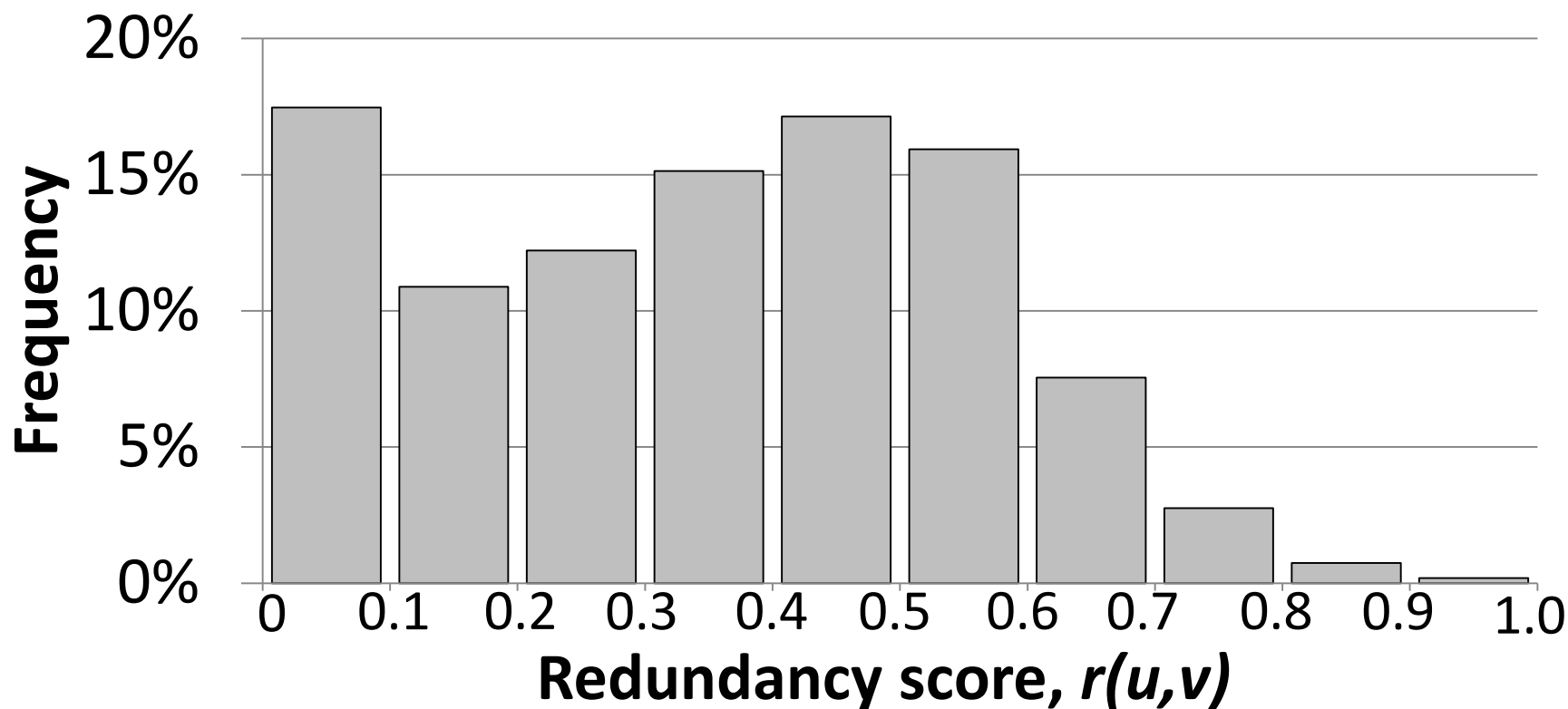
Present documents in both orders. If the top one is always clicked, the documents are duplicate.

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



Score Distribution

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



Classes of Duplication

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

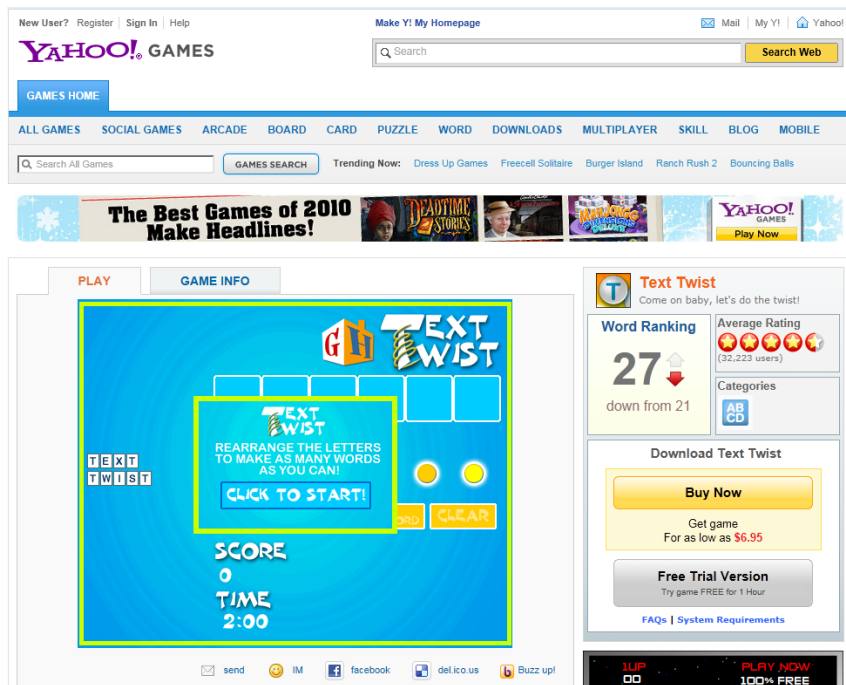
- **Exact duplicates:** Both pages appear identical, perhaps with the exception of ads.
- **Content duplicates:** Both pages provide the same / very related information (for this query), but from different sources
- **Navigational duplicates:** Getting from one page to the other is very easy

Example: Navigational Duplicates



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

Example: Content Duplicates



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

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	<i>any</i>	Yes within	5%	} Navigational duplicates
Left or Right	<i>any</i>	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	<i>any</i>	Yes across	2%	
Neither Relevant	<i>any</i>	<i>any</i>	4%	
<i>other</i>	<i>other</i>	<i>other</i>	3%	

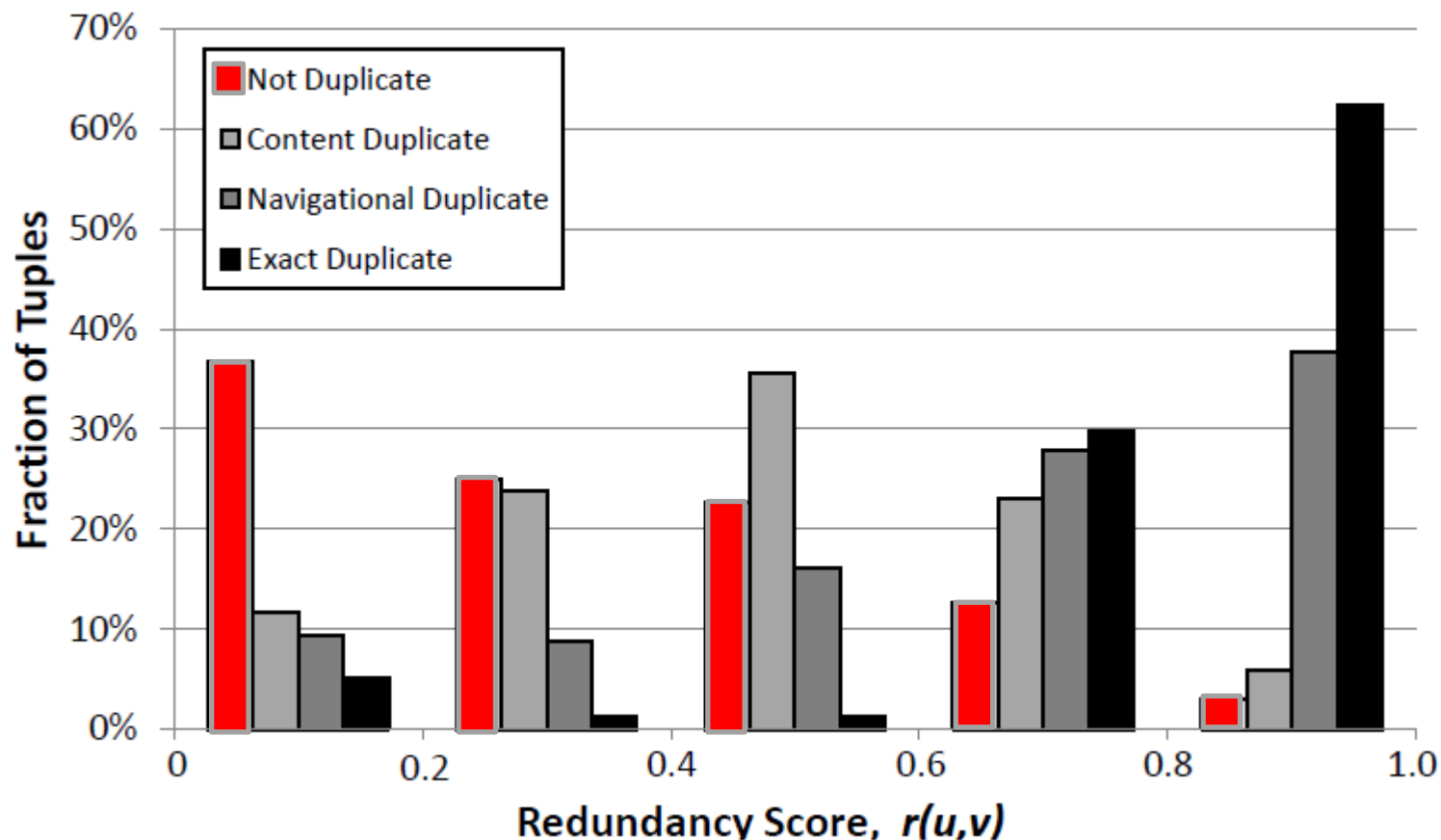
Judging Duplication

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

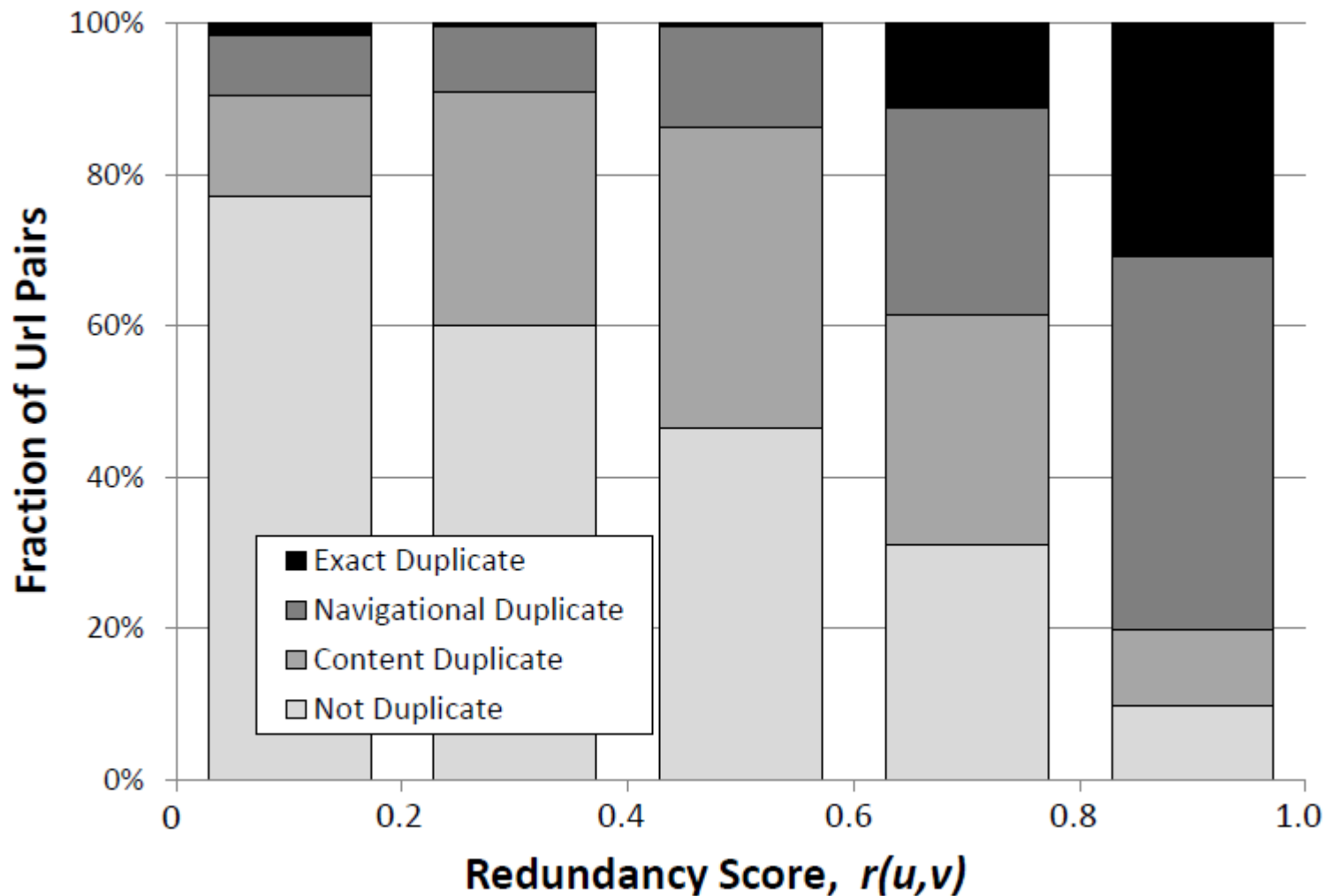
Judgment 2	Judgment 1				
	Ex.	Nav.	Cont.	C_w	Not Dup.
Exact (E)	0	0	0	0	0
Navigational (N)		19	1	3	2
Content (C)			31	27	12
Weak Cont. (C_w)				15	32
Not Duplicate					78

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

Judgments vs Redundancy Score

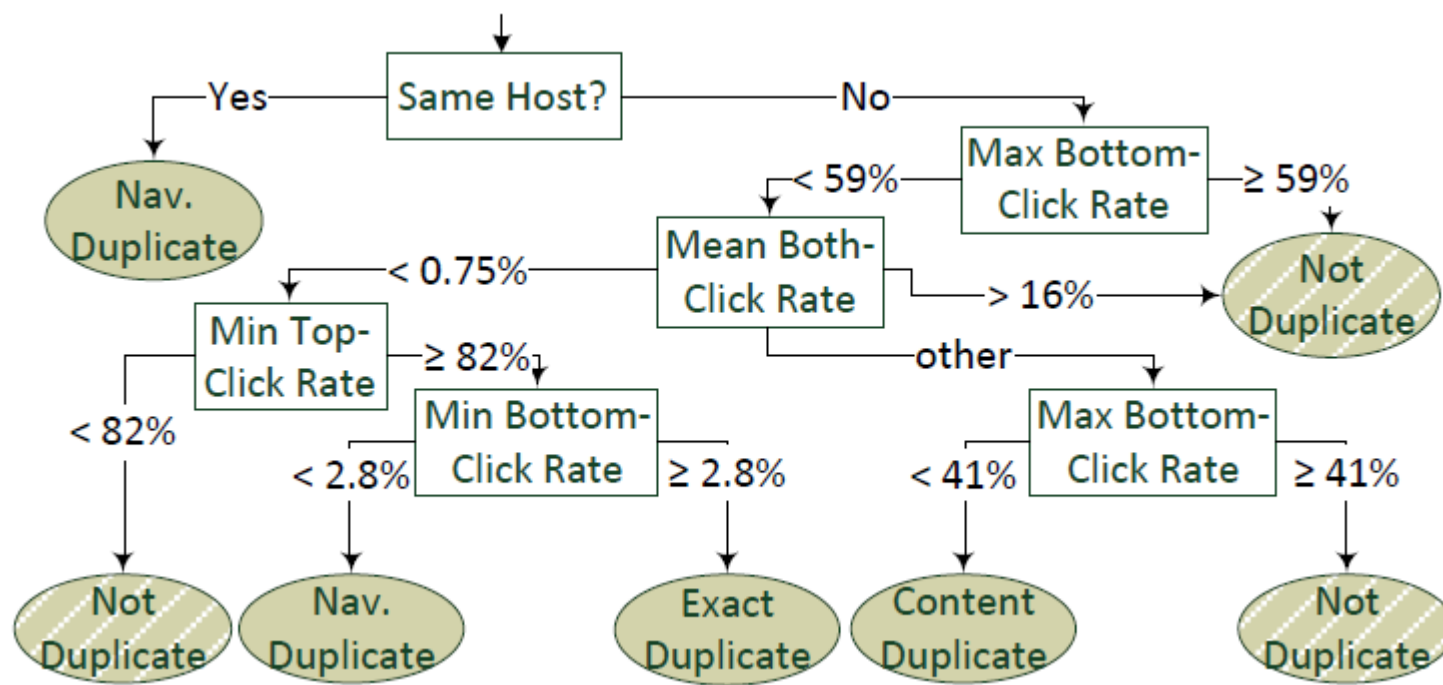


Judgments vs Redundancy Score

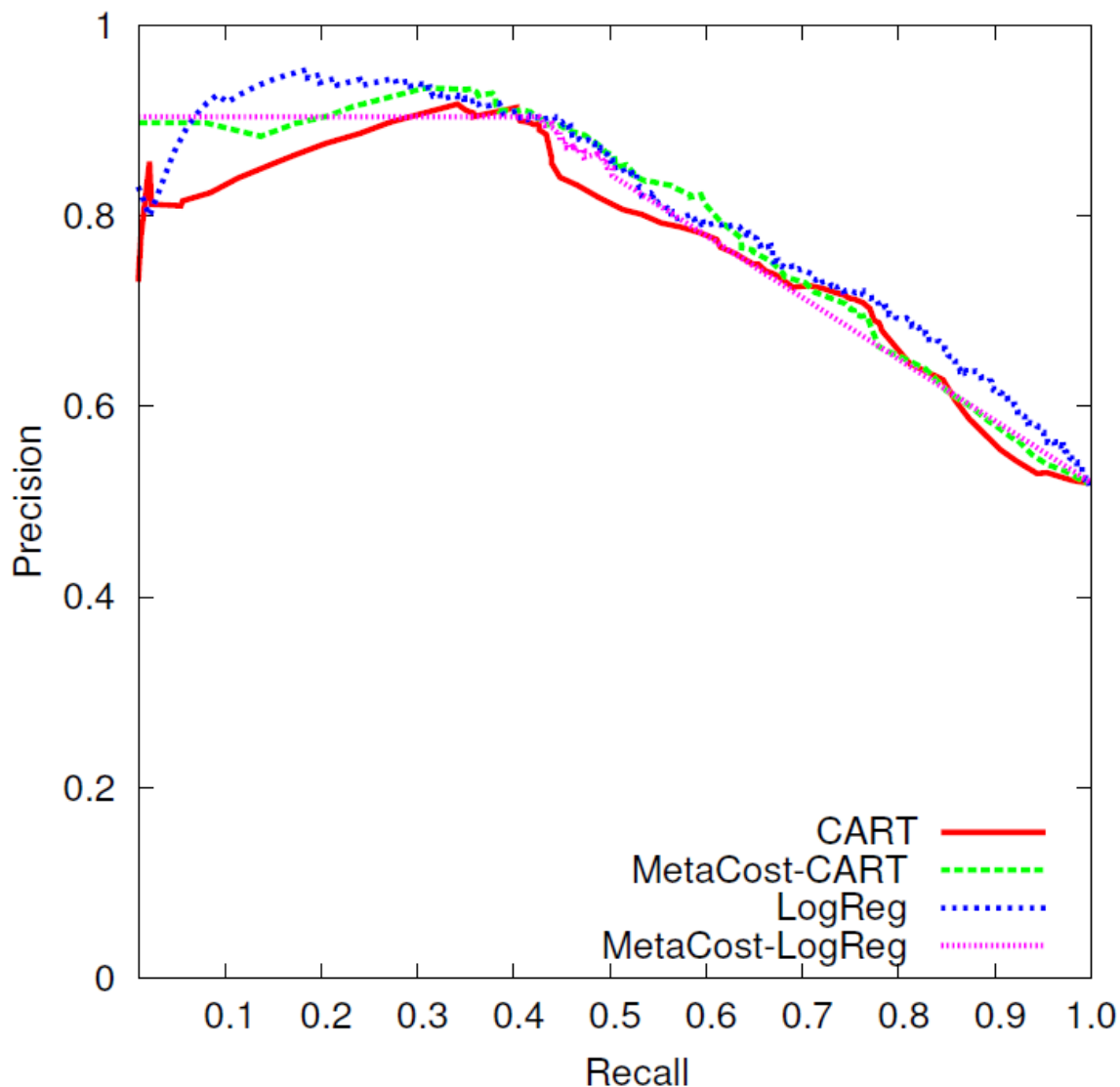


Learning Duplicate Detection

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



Learning Duplicate Detection



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

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.