

Detecting Duplicate Web Documents using Clickthrough Data

Filip Radlinski, Paul N. Bennett, Emine Yilmaz

Microsoft

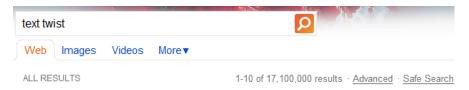
Microsoft[®] Research

Motivation

100

20

- Near-Duplicates and duplicates are common.
- In many cases, showing them isn't helpful.
- Usually identified based on page *content*.
- \succ not sensitive to query.
- We show that clicks give a more context-sensitive duplication signal.



Play Text Twist, download, and read user reviews on Yahoo! Games Download Text Twist, Play Text Twist, and Read Text Twist reviews on Yahoo! Games. games.yahoo.com/game/text-twist · Mark as spam

Text Twist on Yahoo! Games

Text Twist at Yahoo! Games. Read user reviews and play Text Twist as well as dozens of other new and classic games for FREE! games.yahoo.com/console/tx · Mark as spam

Super Text Twist® - Free Online and Downloadable Games and Free ...

Download the full version of Super **Text Twist®** FREE! Play the full version with more features, more levels and better graphics! Download Free Trial or Buy Now Just \$6.99 www.shockwave.com/gamelanding/texttwist.jsp · Mark as spam

TextTwist - MSN Games - Free Online Games

4013 people are playing TextTwist right now. ... Unlimited Play: High Score Tracking 'Twist' Action zone.msn.com/gameplayer/gameplayer.aspx?game=texttwist&instance=default · Mark as spam

Super TextTwist > Play Free Now! | GameHouse

... native speaker of English, the words that are longer than 6 letters are a bit difficult for me, which is why I liked the older version of **Text Twist** better. Still, a very ... www.gamehouse.com/download-games/super-texttwist · Mark as spam

TextTwist - MSN Games - Free Online Games

Twist and turn jumbled letters to make words. Look for patterns and put your vocabulary to the test in the fun, free game TextTwist.

Super TextTwist Online Game > Free! | GameHouse

Online Games Hidden Object Games Zuma Bejeweled 3 Text Twist Diner Dash Jewel Quest Cake Mania Chuzzle

 $www.gamehouse.com/online-games/super-texttwist-online + Mark \ as \ spam$

TextTwist | Play TextTwist Free Online | Games.com

Take the **Text Twist** challenge! Win points when you unscramble the letters to make new words. Play the exciting timed mode or download the full version and relax playing the un ... www.games.com/game/texttwist · Mark as spam



Previous approaches

100

• Detecting Duplication using Content

20

40

– 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.



100

9

× 30 40

0

no

Web Images Videos Sho	pping News Maps More MSN Hotmail	
bing	wsdm 2011 important dates	
Web	Web Videos More V	Click rate $c^{\hat{u}v}$
RELATED SEARCHES	ALL RESULTS 1-10 of 284 results · Advance	
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	home - WSDM2011 WSDM (pronounced "will pn") is the premier, 14 days until coster submission deadline (01/16/2011) IMPORTAN OPCUMENT U www.wsdm2011.org - Caseed Uge Current U Cfp - WSDM2011 4th ACM WSDM Concerner will take place in Hong Kong, during February 9-12, 2011. WSDM Important Dates www.wsdm2011.org/wstm20.0cfp Coste ex page WSDM 2011 Call For Papers Spoonylife 4th ACM WSDM Conference will take place in Hong Kong, during February 9-12, 2011.WSDM	Click rate $c^{u\hat{v}}$
Humorous Calendar Dates Holidays	Important Dates www.spoonylife.org/level-3/wsdm-2011-call-for-papers Cached page	
SEARCH HISTORY Turn on search history to start remembering your searches.	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	
Turn history on	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	
	$a\hat{u}v$	

 $bias_{uv} = \frac{1}{c^{\hat{u}v} + c^{u\hat{v}} + c^{\hat{u}\hat{v}}}$



100

9

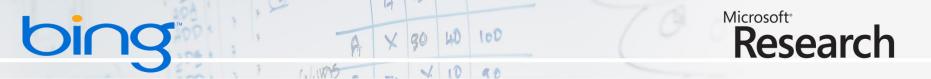
× 30 40

0

no

Web Images Videos Sho	opping News Maps More MSN Hotmail
bing	wsdm 2011 important dates
Web	Web Videos More \checkmark ALL RESULTS 1-10 of 284 results : Advanced \checkmark Click rate $c^{\hat{v}u}$
RELATED SEARCHES	ALL RESULTS 1-10 of 284 results · Advanced
Important Dates to Remember	home - WSDM2011 WSDM (pronounced [with the premier, - 14 days until Coster submission deadline
Important Dates in May Important Dates in History	WSDM (pronounced "wild by") is the premier 14 days uptil foster submission deadline (01/16/2011) IMPOR AN OCCUMENT OF COMPANY OF CALLED AND AND AND AND AND AND AND AND AND AN
Important Dates in U.S. History	<u>cfp - WSDM2011</u> 4th ACM WSDM Correnationwill take placetic More Knest, Auring February 9-12, 2011. WSP
Important Dates in American History	L. 4th ACM WSDM Cor Proceedings of the provided of the provide
Important C alendar Dates	WSDM 2011 Call For Papers Spoonylife 4th ACM WSDM Conference will take place in Hong Kong, during February 9-12, 2011.WSDM
Humorous Calendar Dates Holidays	Important Dates www.spoonylife.org/level-3/wsdm-2011-call-for-papers · Cached page
SEARCH HISTORY	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
Turn on search history to start remembering	original and must not have been published or under review elsewhere. Important Dates ir.ischool.utexas.edu/csdm2011/papers.html · Cached page
your searches. Turn history on	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
	$C^{\hat{\mathcal{V}}\mathcal{U}}$

 $bias_{vu} = \frac{1}{c^{\hat{v}u} + c^{v\hat{u}} + c^{\hat{v}\hat{u}}}$



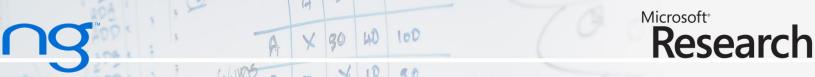
• The redundancy score is the minimum of these two ratios:

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

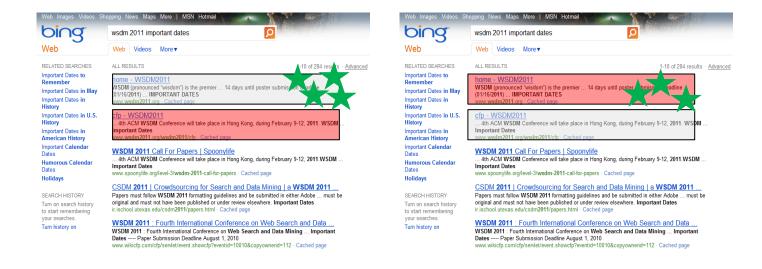
• In words:

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

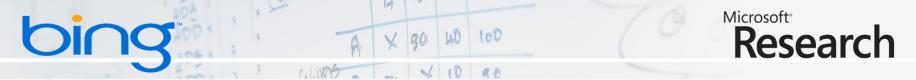
• High presentation bias \leftrightarrow High redundancy score



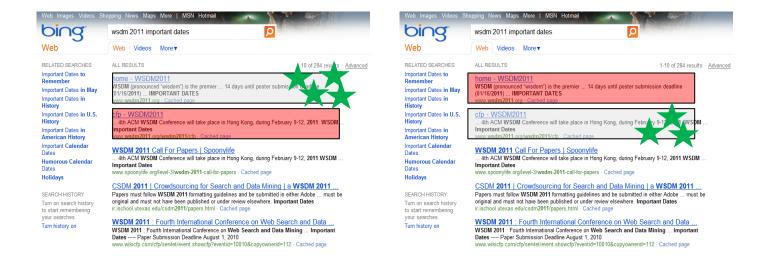
• If in both orders, the top result always gets clicked:



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



 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

Research 100 20 LD

Dates Holidays

- Relationship to Click-Skip and FairPairs bing wsdm 2011 important dates
 - The Click-Skip approach: A clicked document is more relevant than skipped ones above
 - FairPairs:



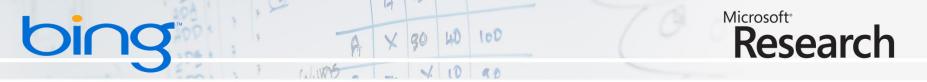
Microsoft[®]

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

100





Microsoft[®]

Research

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

Research

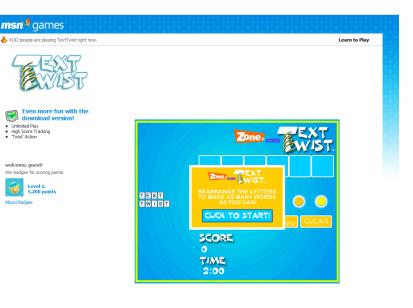
Example: Content Duplicates

100

LD

20

New User? Register Sign In Help		Make Y! My Homepage			🖂 Mail My Y! 🏠 Yahoo!
		Q Search	Search Web		
GAMES HOME					
LL GAMES SOCIAL GAMES	ARCADE BOARD CARD	PUZZLE WORD	DOWNLOADS	MULTIPLAYER SKIL	L BLOG MOBILE
Search All Games	GAMES SEARCH Trendin	ig Now: Dress Up Game	s Freecell Solitaire	Burger Island Ranch Ru	sh 2 Bouncing Balls
The Bes Mak	t Games of 2010 e Headlines!	Contractions of the second			YAHOO! GAMES Play Now
PLAY G	AME INFO		_	Come on bal	ist by, let's do the twist!
	G	EXT WIS	г Г	Word Ranking	Average Rating (32,223 users)
	EXT Ewist			down from 21	Categories
TEXT	REARRANGE THE LETTERS TO MAKE AS MANY WORDS AS YOU CAN!			ad Text Twist	
	CLICK TO START!		3	G	et game low as \$6.95
	0 TJME 2:00			Try game	rial Version FREE for 1 Hour em Requirements
	🖂 send 🥥 IM 💽 face	ebook 📳 del.ico.us	b Buzz up!	· 1UP ·	PLRY NOW



- 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

100

40

9

1

× 30

More Relevant?	Utility?	Navigation?	Freq.	-
Both equally	Identical	-	6%	Exact duplicates
Both equally	any	Yes within	5%	۲ Navigational
Left or Right	any	Yes within	13%	<pre>} Navigational duplicates</pre>
Both equally	Very similar	No	16%	- \ Content
Left or Right	Very similar	No	8%	f duplicates
Both equally	Related	No	4%	"Weak" content
Left or Right	Related	No	15%	duplicates
Different intents	-	No	12%	-
Left or Right	Different	No	12%	Not Duplicate
m Left/Right/Both	any	Yes across	2%	
Neither Relevant	any	any	4%	
other	other	other	3%	J





Judging Duplication

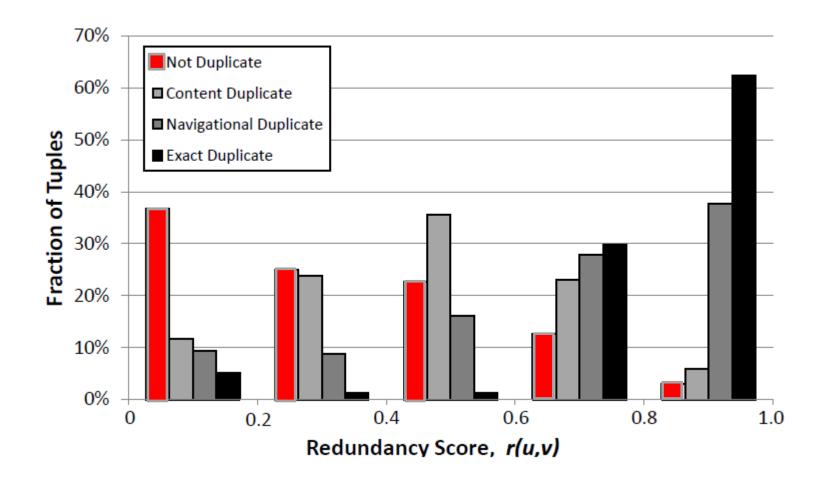
 Inter-judge agreement on a small set tells us its tricky to make these judgments

	Judgment 1				
Judgment 2	Ex.	Nav.	Cont.	$\mathbf{C}_{\boldsymbol{w}}$	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

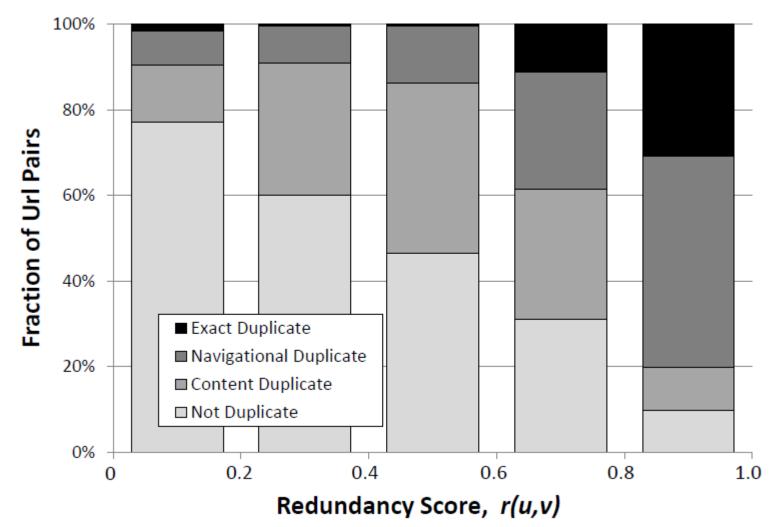


Judgments vs Redundancy Score

100

40

20



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

Microsoft[®]

Research

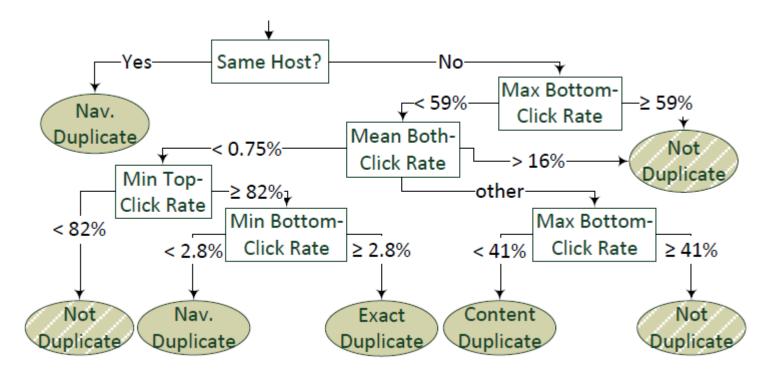
Learning Duplicate Detection

100

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

40

20



Microsoft[®]

Research

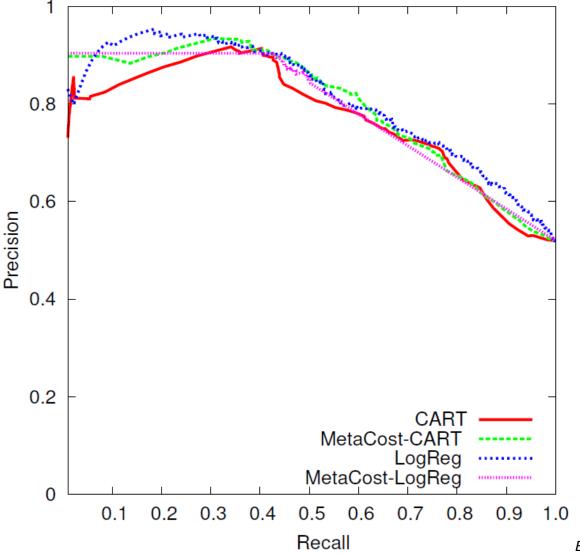
Learning Duplicate Detection

100

40

80

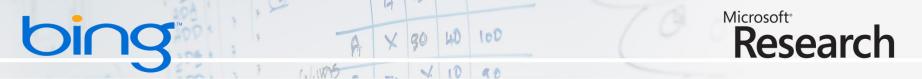
nq



Bennett & Yilmaz, WSDM 2011

Microsoft[®]

Research



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

100

• We proposed a taxonomy of duplication.

20

- Clicks can be used to distinguish among the classes.
- Presentation bias has a limited effect on nonduplicates.
- 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.