Log File Analysis

Maarten de Rijke

Origins of the material

- Joint work with
  - Richard Berendsen, Katja Hofmann, Bouke Huurnink, Edgar Meij, Gilad Mishne, Wouter Weerkamp, Shimon Whiteson
- Slides partially based on material by
  - Andrei Broder, Jim Jansen, Danny Levinson, Daniel Rose, Fabrizio Silvestri

2

UNIVERSITEIT VAN AMSTERDAM

AOL-user-ct-collection — less — 116×53

```
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:13
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:33
711391  can not sleep with snoring husband      2006-03-01 01:24:00
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:33:06     1       http://www.epinions.com
711391  jackie zeaman nude      2006-03-01 15:26:27
711391  jackie zeman nude       2006-03-01 15:26:38
711391  strange cosmos  2006-03-01 16:07:15     1       http://www.strangecosmos.com
711391  mansfield first assembly        2006-03-01 16:09:20     1       http://www.mansfieldfirstassembly.org
711391  mansfield first assembly        2006-03-01 16:09:20     3       http://netministries.org
711391  reverend harry myers    2006-03-01 16:10:07
711391  reverend harry myers    2006-03-01 16:10:30
711391  national enquirer       2006-03-01 17:13:14     1       http://www.nationalenquirer.com
711391  how to kill mockingbirds        2006-03-01 17:18:11
711391  how to kill mockingbirds        2006-03-01 17:18:33
711391  how to kill annoying birds in your yards        2006-03-01 17:18:58
711391  how to kill annoying birds in your yards        2006-03-01 17:19:53     2       http://www.sortprice.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     3       http://shopping.msn.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     10      http://www.bergen.org
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:24:35     15      http://www.saferbrand.com
711391  how do i get mocking birds out of my yard       2006-03-01 17:27:17
711391  how do i get mockingbirds out of my yard        2006-03-01 17:27:36     9       http://www.asri.org
711391  how do i get mockingbirds out of my yard        2006-03-01 17:30:14
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     3       http://www.bird-x.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     1       http://forums2.gardenweb.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     10      http://www.birding.com
711391  mansfield first assembly        2006-03-01 18:31:36     3       http://netministries.org
711391  beth moore      2006-03-01 19:42:41     1       http://www.lproof.org
711391  judy baker ministries   2006-03-01 19:49:03     2       http://www.embracinggrace.com
711391  god will fulfill your hearts desires    2006-03-01 19:59:06     10      http://www.pureintimacy.org
711391  online friendships can be very special  2006-03-01 23:09:37
711391  online friendships can be very special  2006-03-01 23:09:57
711391  online friendships      2006-03-01 23:10:24
711391  cypress fairbanks isd   2006-03-02 07:56:53     1       http://www.cfisd.net
711391  people are not always how they seem over the internet   2006-03-02 08:31:51
711391  friends online can be different in person       2006-03-02 08:32:42
711391  friends online can be different in person       2006-03-02 08:33:04     13      http://www.salon.com
711391  boston butts    2006-03-02 09:47:36
711391  community christian church houston tx   2006-03-02 16:07:53
```

```
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:33:06     1       http://www.epinions.com
711391  jackie zeaman nude      2006-03-01 15:26:27
711391  jackie zeman nude       2006-03-01 15:26:38
711391  strange cosmos  2006-03-01 16:07:15     1       http://www.strangecosmos.com
711391  mansfield first assembly        2006-03-01 16:09:20     1       http://www.mansfieldfirstassembly.org
711391  mansfield first assembly        2006-03-01 16:09:20     3       http://netministries.org
711391  reverend harry myers    2006-03-01 16:10:07
711391  reverend harry myers    2006-03-01 16:10:30
711391  national enquirer       2006-03-01 17:13:14     1       http://www.nationalenquirer.com
711391  how to kill mockingbirds        2006-03-01 17:18:11
711391  how to kill mockingbirds        2006-03-01 17:18:33
711391  how to kill annoying birds in your yards        2006-03-01 17:18:58
711391  how to kill annoying birds in your yards        2006-03-01 17:19:53     2       http://www.sortprice.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     3       http://shopping.msn.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     10      http://www.bergen.org
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:24:35     15      http://www.saferbrand.com
711391  how do i get mocking birds out of my yard       2006-03-01 17:27:17
711391  how do i get mockingbirds out of my yard        2006-03-01 17:27:36     9       http://www.asri.org
711391  how do i get mockingbirds out of my yard        2006-03-01 17:30:14
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     3       http://www.bird-x.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     1       http://forums2.gardenweb.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     10      http://www.birding.com
711391  mansfield first assembly        2006-03-01 18:31:36     3       http://netministries.org
711391  beth moore      2006-03-01 19:42:41     1       http://www.lproof.org
711391  judy baker ministries   2006-03-01 19:49:03     2       http://www.embracinggrace.com
711391  god will fulfill your hearts desires    2006-03-01 19:59:06     10      http://www.pureintimacy.org
711391  online friendships can be very special  2006-03-01 23:09:37
711391  online friendships can be very special  2006-03-01 23:09:57
711391  online friendships      2006-03-01 23:10:24
711391  cypress fairbanks isd   2006-03-02 07:56:53     1       http://www.cfisd.net
711391  people are not always how they seem over the internet   2006-03-02 08:31:51
711391  friends online can be different in person       2006-03-02 08:32:42
711391  friends online can be different in person       2006-03-02 08:33:04     13      http://www.salon.com
711391  boston butts    2006-03-02 09:47:36
711391  community christian church houston tx   2006-03-02 16:07:53
711391  gay churches in houston tx      2006-03-02 16:08:23
711391  community gospel church in houston tx   2006-03-02 16:08:45     2       http://www.communitygospel.org
711391  houston tx is one hot place     2006-03-02 18:04:44
711391  houston tx is one hot place to live     2006-03-02 18:04:55     9       http://travel.yahoo.com
711391  houston tx is one hot place to live     2006-03-02 18:16:05     1       http://www.houston-texas-online.com
711391  texas hill country and sights around san antonio tx     2006-03-02 18:19:00     5       http://www.answers.c
om
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:04
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:30     1       http://www.askdoctrish.com
711391  strange cosmos  2006-03-02 19:29:31     1       http://www.strangecosmos.com
711391  white hard dry skin on face     2006-03-02 20:31:29
711391  white hard dry skin on face     2006-03-02 20:32:24
:
```

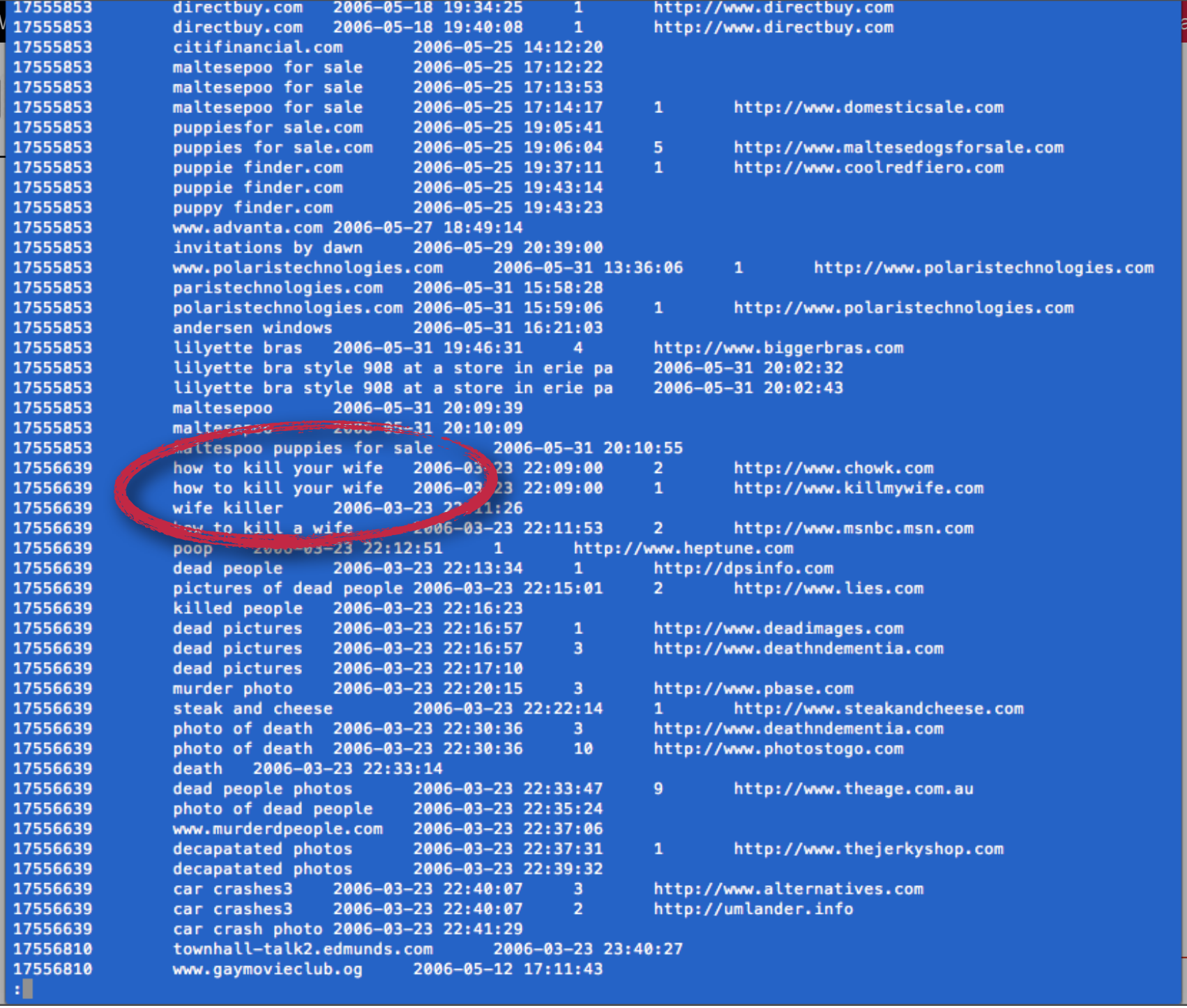
Another example

4

UNIVERSITEIT VAN AMSTERDAM

AOL-user-ct-collection — less — 116×53

```
17555853        www.addresses.com       2006-05-18 15:39:20     1       http://www.addresses.com
17555853        www.addresses.com       2006-05-18 15:39:20     1       http://www.addresses.com
17555853        www.addresses.com       2006-05-18 15:39:20     1       http://www.addresses.com
17555853        directbuy.com   2006-05-18 19:34:25     1       http://www.directbuy.com
17555853        directbuy.com   2006-05-18 19:40:08     1       http://www.directbuy.com
17555853        citifinancial.com       2006-05-25 14:12:20
17555853        maltesepoo for sale     2006-05-25 17:12:22
17555853        maltesepoo for sale     2006-05-25 17:13:53
17555853        maltesepoo for sale     2006-05-25 17:14:17     1       http://www.domesticsale.com
17555853        puppiesfor sale.com     2006-05-25 19:05:41
17555853        puppies for sale.com    2006-05-25 19:06:04     5       http://www.maltesedogsforsale.com
17555853        puppie finder.com       2006-05-25 19:37:11     1       http://www.coolredfiero.com
17555853        puppie finder.com       2006-05-25 19:43:14
17555853        puppy finder.com        2006-05-25 19:43:23
17555853        www.advanta.com 2006-05-27 18:49:14
17555853        invitations by dawn     2006-05-29 20:39:00
17555853        www.polaristechnologies.com     2006-05-31 13:36:06     1       http://www.polaristechnologies.com
17555853        paristechnologies.com   2006-05-31 15:58:28
17555853        polaristechnologies.com 2006-05-31 15:59:06     1       http://www.polaristechnologies.com
17555853        andersen windows        2006-05-31 16:21:03
17555853        lilyette bras   2006-05-31 19:46:31     4       http://www.biggerbras.com
17555853        lilyette bra style 908 at a store in erie pa    2006-05-31 20:02:32
17555853        lilyette bra style 908 at a store in erie pa    2006-05-31 20:02:43
17555853        maltesepoo      2006-05-31 20:09:39
17555853        maltesepoo      2006-05-31 20:10:09
17555853        maltespoo puppies for sale      2006-05-31 20:10:55
17556639        how to kill your wife   2006-03-23 22:09:00     2       http://www.chowk.com
17556639        how to kill your wife   2006-03-23 22:09:00     1       http://www.killmywife.com
17556639        wife killer     2006-03-23 22:11:26
17556639        how to kill a wife      2006-03-23 22:11:53     2       http://www.msnbc.msn.com
17556639        poop    2006-03-23 22:12:51     1       http://www.heptune.com
17556639        dead people     2006-03-23 22:13:34     1       http://dpsinfo.com
17556639        pictures of dead people 2006-03-23 22:15:01     2       http://www.lies.com
17556639        killed people   2006-03-23 22:16:23
17556639        dead pictures   2006-03-23 22:16:57     1       http://www.deadimages.com
17556639        dead pictures   2006-03-23 22:16:57     3       http://www.deathndementia.com
17556639        dead pictures   2006-03-23 22:17:10
17556639        murder photo    2006-03-23 22:20:15     3       http://www.pbase.com
17556639        steak and cheese        2006-03-23 22:22:14     1       http://www.steakandcheese.com
17556639        photo of death  2006-03-23 22:30:36     3       http://www.deathndementia.com
```

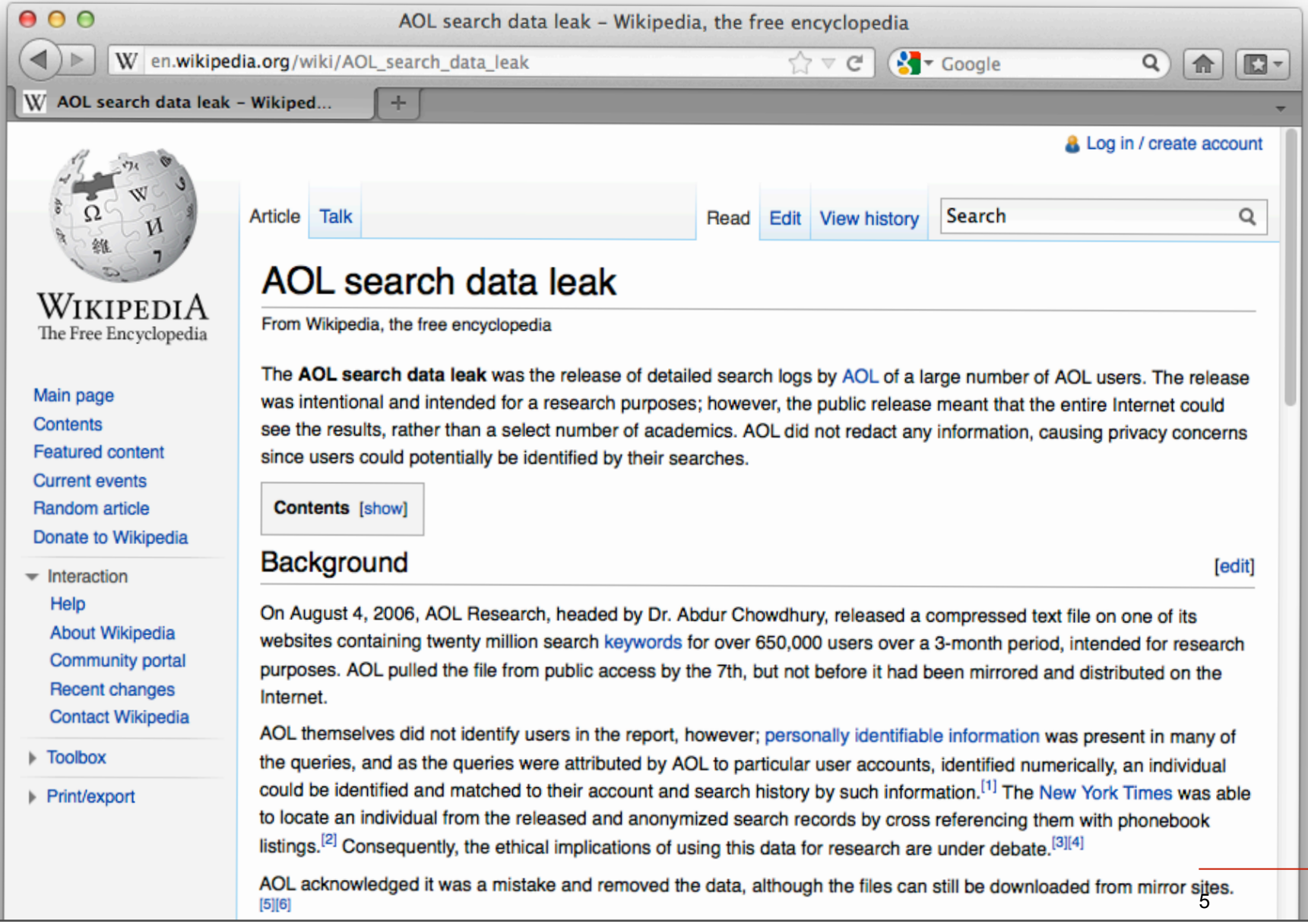


```
17555853        directbuy.com   2006-05-18 19:34:25     1       http://www.directbuy.com
17555853        directbuy.com   2006-05-18 19:40:08     1       http://www.directbuy.com
17555853        citifinancial.com       2006-05-25 14:12:20
17555853        maltesepoo for sale     2006-05-25 17:12:22
17555853        maltesepoo for sale     2006-05-25 17:13:53
17555853        maltesepoo for sale     2006-05-25 17:14:17     1       http://www.domesticsale.com
17555853        puppiesfor sale.com     2006-05-25 19:05:41
17555853        puppies for sale.com    2006-05-25 19:06:04     5       http://www.maltesedogsforsale.com
17555853        puppie finder.com       2006-05-25 19:37:11     1       http://www.coolredfiero.com
17555853        puppie finder.com       2006-05-25 19:43:14
17555853        puppy finder.com        2006-05-25 19:43:23
17555853        www.advanta.com 2006-05-27 18:49:14
17555853        invitations by dawn     2006-05-29 20:39:00
17555853        www.polaristechnologies.com     2006-05-31 13:36:06     1       http://www.polaristechnologies.com
17555853        paristechnologies.com   2006-05-31 15:58:28
17555853        polaristechnologies.com 2006-05-31 15:59:06     1       http://www.polaristechnologies.com
17555853        andersen windows        2006-05-31 16:21:03
17555853        lilyette bras   2006-05-31 19:46:31     4       http://www.biggerbras.com
17555853        lilyette bra style 908 at a store in erie pa    2006-05-31 20:02:32
17555853        lilyette bra style 908 at a store in erie pa    2006-05-31 20:02:43
17555853        maltesepoo      2006-05-31 20:09:39
17555853        malte[illegible]  2006-05-31 20:10:09
17555853        maltespoo puppies for sale      2006-05-31 20:10:55
17556639        how to kill your wife   2006-03-23 22:09:00     2       http://www.chowk.com
17556639        how to kill your wife   2006-03-23 22:09:00     1       http://www.killmywife.com
17556639        wife killer     2006-03-23 22:11:26
17556639        how to kill a wife      2006-03-23 22:11:53     2       http://www.msnbc.msn.com
17556639        poop    2006-03-23 22:12:51     1       http://www.heptune.com
17556639        dead people     2006-03-23 22:13:34     1       http://dpsinfo.com
17556639        pictures of dead people 2006-03-23 22:15:01     2       http://www.lies.com
17556639        killed people   2006-03-23 22:16:23
17556639        dead pictures   2006-03-23 22:16:57     1       http://www.deadimages.com
17556639        dead pictures   2006-03-23 22:16:57     3       http://www.deathndementia.com
17556639        dead pictures   2006-03-23 22:17:10
17556639        murder photo    2006-03-23 22:20:15     3       http://www.pbase.com
17556639        steak and cheese        2006-03-23 22:22:14     1       http://www.steakandcheese.com
17556639        photo of death  2006-03-23 22:30:36     3       http://www.deathndementia.com
17556639        photo of death  2006-03-23 22:30:36     10      http://www.photostogo.com
17556639        death   2006-03-23 22:33:14
17556639        dead people photos      2006-03-23 22:33:47     9       http://www.theage.com.au
17556639        photo of dead people    2006-03-23 22:35:24
17556639        www.murderdpeople.com   2006-03-23 22:37:06
17556639        decapatated photos      2006-03-23 22:37:31     1       http://www.thejerkyshop.com
17556639        decapatated photos      2006-03-23 22:39:32
17556639        car crashes3    2006-03-23 22:40:07     3       http://www.alternatives.com
17556639        car crashes3    2006-03-23 22:40:07     2       http://umlander.info
17556639        car crash photo 2006-03-23 22:41:29
17556810        townhall-talk2.edmunds.com      2006-03-23 23:40:27
17556810        www.gaymovieclub.og     2006-05-12 17:11:43
:
```

UNIVERSITEIT VAN AMSTERDAM

5

UNIVERSITEIT VAN AMSTERDAM



5

CENSORSHIP

just brighter

HOT TOPICS EDITOR'S PICKS APPLE FACEBOOK SOPA ANDROID CRUNCHIES AWARDS

Congress is considering legislation (SOPA and PIPA) that could alter the basic rules of the road o
fight foreign piracy of movies, music, and other entertainment content. Learn more. Read our full

Comment 3 Like 14 Tweet 3 Share +1 1

AOL Proudly Releases Massive Amounts of Private Data

MICHAEL ARRINGTON | Sunday, August 6th, 2006 | 3 Comments

Yet Another Update: AOL: "This was a screw up"

Further Update: Sometime after 7 pm the download link went down as well, but there is at least one mirror site. AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.

Update: Sometime around 7 pm PST on Sunday, the AOL site referred to below was taken down. The direct link to the data is still live. A cached copy of the page is here.

AOL must have missed the uproar over the DOJ's demand for "anonymized" search data last year that caused all sorts of pain for Microsoft and Google. That's the only way to explain their release of data that includes 20 million web queries from 650,000 AOL users.

The data includes all searches from those users for a three month period this year, as well as whether

...files can still be downloaded from mirror sites.

5


AOL Search Data Shows Users Planning to commit Murder. « Plenty of fish blog
http://plentyoffish.wordpress.com/2006/08/07/aol-search-data-show
Reader
Google
Plenty of fish blog
Adapt or die - by Markus Frind CEO of Plentyoffish.com
« AOL Search Data Shows Myspace growing via SEO SPAM.
Bebo And Myspace.com Reality Check based on $436 test »
AOL Search Data Shows Users Planning to commit Murder.
*** Update**** Monday July 7th 7 PM PST
Users in the comments are pissed off at the idea that people can be arrested for planning a crime like murder, calling it minority report like. I ask you why is it that americans have no problems arresting people that are planning or researching how to conduct terrorist attacks? Yet if a person plans on killing his wife that is ok, until he actually does it? How many people do you have to plan on killing before its ok for a company like AOL to hand your records over to the government? I am not taking sides, I'm just pointing out the obvious double standard. This story will open a can of worms, and will decide just how private your data online really is.
http://research.aol.com released a list of 20 million + searches by 500,000 AOL users. Contained in this list are social security numbers, credit cards and other personal information. There are some truly scary things in this database.
There are hundreds of searches from people looking to kill themselves and even more scary are searches from users that seem to be looking to commit murder.
Check out the search history for user 17556639, most recent search

What are we looking at?

- It is all about **behavior**: observable activities of a person, a team, a system, …
  - Something that we can detect and record
  - Actions with some purpose
  - Responses to stimuli
- **Not**
  - Affective, cognitive, situational aspects
- Not just isolated snapshots but **trace** data
  - People conducting sequences of activities
  - Repeat activities
  - Development of behavior over time
  - Repeat searches, sessions, personalization

Behavior	Description
View results	Interaction in which user viewed or scrolled result pages
With scrolling	*User scrolled result page*
Without scrolling	*User did not scroll result page*
…	
Selection	Interaction in which user makes a selection in the result pages
Click URL	*User clicked on a URL*
Next in Set of Result Pages	*User moved to the next result page*
…	

UNIVERSITEIT VAN AMSTERDAM

Some reflections

7

Some reflections

- Data collection advantages of log data
  - **Scale**: not a limiting factor as in lab user studies
  - **Power**: large sample size for inference
  - **Scope**: allows for study of a range of interactions in a multi-variable context
  - **Location**: can be collected in distributed environments
  - **Duration**: can be collected over extended periods

7

Some reflections

- Data collection advantages of log data
  - **Scale**: not a limiting factor as in lab user studies
  - **Power**: large sample size for inference
  - **Scope**: allows for study of a range of interactions in a multi-variable context
  - **Location**: can be collected in distributed environments
  - **Duration**: can be collected over extended periods
- Allows data collection without directly interfering with users
  - No observer effect, No observer bias, …
  - Might still happen for data analysis

7

Some reflections

- Data collection advantages of log data
  - **Scale**: not a limiting factor as in lab user studies
  - **Power**: large sample size for inference
  - **Scope**: allows for study of a range of interactions in a multi-variable context
  - **Location**: can be collected in distributed environments
  - **Duration**: can be collected over extended periods
- Allows data collection without directly interfering with users
  - No observer effect, No observer bias, …
  - Might still happen for data analysis
- Issues
  - Abstraction, selection, reduction
  - Context

7

Why are logs interesting?

8

Why are logs interesting?

- People watching

Why are logs interesting?

- People watching
- Records of interactions between humans and information retrieval engines

Why are logs interesting?

- People watching
- Records of interactions between humans and information retrieval engines

- Why bother
  - Understand interaction behavior
  - Understand engine usage
  - Optimize engine usage

Why are logs interesting?

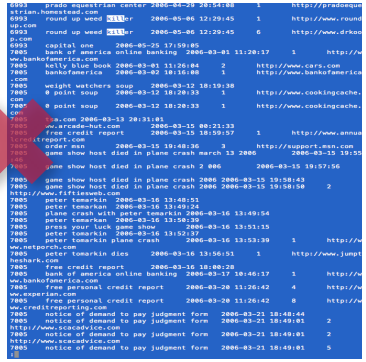
- People watching
- Records of interactions between humans and information retrieval engines

- Why bother
  - Understand interaction behavior
  - Understand engine usage
  - Optimize engine usage

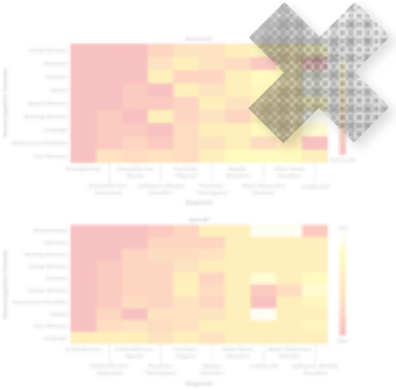
- And what would you do with it?
  - Improve system design
  - Improve models of user behavior
  - Advance searching assistance
  - Revise evaluation methodology
  - …

UNIVERSITEIT VAN AMSTERDAM

What



Examples



Analysis

Uses

What to do?

Wrap-up

Examples

- Long history of transaction log analysis
  - Early studies of the logs created by users of library online public access catalog systems (Peters, 1993)
  - Early use Web search engine logs (Jansen and Pooch, 2001)
  - General overview (Silvestri, 2010)
  - More specialized search engines and their transaction logs
    - A few examples
- Three frequently used units of analysis
  - the **session**, the **query**, the **term**
  - Definitions vary across studies

Peters, 1993

- Overview of “the first 25 years of Transaction Log Analysis”
  - Instructive to get a feel for diversity, questions and goals
- “A pioneer mentality has pervaded the first quarter century of transaction log analysis. Researchers have charged into this new territory created by the development of automated IR systems in general and transaction logging facilities in particular. … Researhcers are still exploring exploring the boundaries and interior terrain of the area opened up by automated IR systems with their transaction logging capabilities.”
- T.A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2):41–66, 1993.

UNIVERSITEIT VAN AMSTERDA

Peters, 1993

- Overview of "the
  - Instructive to
- "A pioneer menta
  transaction log ar
  territory created b
  general and trans
  Researhcers are
  terrain of the area
  transaction loggir
- T.A. Peters. The l
  Library Hi Tech, 1

ew
in
terior
eir
alysis.

SIDEBAR 2: TRANSACTION LOG ANALYSES SORTED BY SYSTEM (1980 TO PRESENT)

SYSTEM	SITE	YEAR PUBLISHED	PRINCIPAL AUTHOR
ADONIS	University College London	1991	Richardson
BIS	U. of North Carolina	1986	Chang
BIS	U. of North Carolina	1987	Bennett, D. B.
BIS	U. of NC—Chapel Hill	1988	Stockton
BIS	U. of NC—Chapel Hill	1989	Weakley
BIS	North Carolina State U.	1990	Taylor
BIS	North Carolina State U.	1991	Hunter
BLEND	U. of Loughborough	1987	Pullinger
BRS	U. of Wisconsin-Stout	1984	Trzebiatowski
BRS	Medline	1986	Kirby
BRS	U. of IL—Urbana-Champaign	1989	Mischo
CATLINE	Nat. Library of Medicine	1984	Tolle
COMPUSERVE	Grolier's Ac Am Encyclopedia	1987	Marchionini
DIALOG	Magazine ASAP	1990	Tenopir
DIALOG	ERIC	1990	Heith-Yee
DIALOG	Rutgers U.	1990	Saracevic
DIALOG	Rutgers U.	1991	Saracevic
DRA	Drew U.	1993	Snelson
GEAC	U. of Sussex	1986	Young
GEAC	U. of Ottawa	1988	Holmes
GRATEFULMED	Johns Hopkins U.	1989	Cahan
ILLINET	Illinois	1991	Connell
ILS	U. of Maryland	1984	Freiburger
INNOPAC	U. of Nevada at Reno	1991	Zink
INNOPAC	Adelphi U.	1992	Ballard
LCS	Ohio State U.	1981	Norden
LCS	Ohio State U.	1982	Borgman
LCS	Ohio State U.	1983	Borgman
LCS	Ohio State U.	1986	Janosky
LCS	U. of IL—Chicago	1989	Wiberley
LCS	Ohio State U.	1992	Helmick
LIAS	Penn State U.	1991	Kalin
LS/2000	U. of Newcastle	1988	Barber
LS/2000	Hampshire College	1985	Hildreth
LS/2000	Newcastle U.	1991	Jeffreys
LURES	Sheffield U.	1983	Meadow
MEDLINE	Nat. Library of Medicine	1981	Penniman
MEDLINE	Case Western Reserve	1984	Woelfl
MEDLINE	Medical College of Pennsylvania	1988	Miller
MEDLINE	U. of Southern California	1992	Nelson
MELVYL	U. of California	1981	Larson
MELVYL	U. of California	1982	Berger
MELVYL	U. of California	1983	Larson
MELVYL	U. of California	1984	Lawrence
MELVYL	U. of California	1986	Larson
MELVYL	U. of California	1989	Larson
MELVYL	U. of California	1991	Larson
MSUS/PALS	Minnesota	1990	Flaherty
MSUS/PALS	Minnesota	1992	Hastuft
multiple	multiple	1983	Kern-Simirenko
multiple	multiple	1983	Tolle
multiple	multiple	1983	Larson
multiple	Nat. Library of Medicine	1985	Tolle

SIDEBAR 2: TRANSACTION LOG ANALYSES SORTED BY SYSTEM (1980 TO PRESENT)

SYSTEM	SITE	YEAR PUBLISHED	PRINCIPAL AUTHOR
ADONIS	University College London	1991	Richardson
BIS	U. of North Carolina	1986	Chang
BIS	U. of North Carolina	1987	Bennett, D. B.
BIS	U. of NC—Chapel Hill	1988	Stockton
BIS	U. of NC—Chapel Hill	1989	Weakley
BIS	North Carolina State U.	1990	Taylor
BIS	North Carolina State U.	1991	Hunter
BLEND	U. of Loughborough	1987	Pullinger
BRS	U. of Wisconsin-Stout	1984	Trzebiatowski
BRS	Medline	1986	Kirby
BRS	U. of IL—Urbana-Champaign	1989	Mischo
CATLINE	Nat. Library of Medicine	1984	Tolle
COMPUSERVE	Grolier's Ac Am Encyclopedia	1987	Marchionini
DIALOG	Magazine ASAP	1990	Tenopir
DIALOG	ERIC	1990	Hseih-Yee
DIALOG	Rutgers U.	1990	Saracevic
DIALOG	Rutgers U.	1991	Saracevic
DRA	Drew U.	1993	Snelson
GEAC	U. of Sussex	1986	Young
GEAC	U. of Ottawa	1988	Holmes
GRATEFULMED	Johns Hopkins U.	1989	Cahan
ILLINET	Illinois	1991	Connell
ILS	U. of Maryland	1984	Freiburger
INNOPAC	U. of Nevada at Reno	1991	Zink
INNOPAC	Adelphi U.	1992	Ballard
LCS	Ohio State U.	1981	Norden
LCS	Ohio State U.	1982	Borgman
LCS	Ohio State U.	1983	Borgman
LCS	Ohio State U.	1986	Janosky
LCS	U. of IL—Chicago	1989	Wiberley

Jansen and Pooch, 2001

Jansen and Pooch, 2001

- Survey and comparison of log studies
  - Early web log studies
    - Fireball (German Web search engine, late 1990s)
    - Excite (Web serach engine, late 1990s)
    - Altavista (Web search engine, late 1990s)
  - Traditional IR systems
    - Hsieh-Yee (1993), Koenemann and Belkin (1996), Siegfried, Bates and Wilde (1993).
  - OPAC type systems
    - Millsap and Ferl (1993), Peters (1989), Wallace (1993)

TABLE 1. Comparison of Web-user studies.

Category	Fireball study	Excite study	Alta Vista study
Period of data collection	31 days 1–31 July 98	Portion of 1 day 10 March 1997	43 days 2 Aug–13 Sept 98
Web IR system	Fireball search engine	Excite search engine	Alta Vista search engine
Document collection size at time of data collection (approx.)	3 million Web sites	30 to 50 million Web sites	100 million documents
Number of queries in data set	16,252,902	54,573	993,208,159
Session length (number of queries in session); sd = standard deviation	Not reported	Mean = 1.6, sd = 0.69 One: 67% (36,564) Two: 19% (10,391) Three: 7% (3,820) Four: 3% (1,637) Over Four: 4% (2,183)	Mean = 2.02, sd = 123.4* One: 77.6% (221,527,914) Two: 13.5% (38,539,006) Three: 4.4% (12,560,861) More Than Three: 4.5% (12,846,335) *the large sd may be due to softbots
Query length (number of terms in query); sd = standard deviation	Mean = 1.66 sd = 0.70 Zero: Not reported. One: 54.59% (8,873,001) Two: 30.80% (5,005,653) Three: 10.36% (1,683,129) More Than Three: 4% (691,119)	Mean = 2.21 sd = 1.05 Zero: 5.02% (2,584) One: 30.81% (15,854) Two: 31.46% (16,191) Three: 17.96% (9,242) More than three: 15% (8,186)	Mean = 2.35 sd = 1.74 Zero: 20.6% (204,600,881) One: 25.8% (256,247,705) Two: 26.0% (258,243,121) Three: 15.0% (148,981,224) More Than Three: 12.6% (125,144,228)
Use of Boolean (queries containing Boolean operators)	2.55% (414,461) *maximum possible number based on data provided	8.54% (4,661)	Not reported *see use of modifiers
Failure rate (improperly structured queries)	Not reported	10% (5,457)	Not reported
Use of modifiers (e.g., +, −, NEAR, etc.) (queries containing a modifier)	25.3% (4,111,843)	9% (4,776)	20.4% (202,614,464)* *Includes Boolean operators
Number of relevant documents viewed in a session	10 or Less: 59.51% (9,621,347) More than 10: 40.47 (6,545,887)	10 or less: 58% (31,652) More than 10: 42% (14,735)	10 or less: 85.2% More than 10: 14.8% *Numbers not reported and not calculable based on data provided.

TABLE 1. Comparison of Web-user studies.

viewed in a session	More than 10: 40.47 (6,545,887)	More than 10: 42% (14,735)	More than 10: 14.8%

*Numbers not reported and not calculable based on data provided.

TABLE 2. Comparison of three traditional IR-user studies.

Category	Koenemann & Belkin study	Hsieh-Yee study	Siegfried, Bates, & Wilde Study
Number of users and experience level	64 novice	30 novice and 32 experts	21 novice
Document collection utilized	74,520 articles from TREC	ERIC database	6 databases on humanities topics
IR system utilized	INQUERY	DIALOG	DIALOG
Session length (number of queries per user per session); sd = standard deviation	Mean = 7 Median = 8.2 *cannot determine sd from data provided	Not reported	Mean = 16.6 sd = 13.5
Query length (number of terms per query); sd = standard deviation	Mean = 6.4 sd = 4.2 *Terms in quotes counted as one term.	Mean for novice = 8.77 Mean for experts = 7.28	62.5% (2,563) of queries were one term 37.5% (1,538) of queries were two terms or more
Use of Boolean (number of queries containing Boolean operators)	Not reported	Not reported	36.8% (1,509) of queries contained one or more Boolean operator
Use of advanced features (number of queries containing advanced options)	Not reported	Mean for novice = 8.80 Mean for experts = 15.69	20.3% (832) of the queries contained one or more advanced feature *does not include use of Boolean operators
Failure rate (number of queries improperly formatted)	Not reported	Not reported	17% (697) of the queries contained a formatting error
Number of relevant documents viewed Per session	Not reported	Mean for novice = 10.31 Mean for experts = 28.72	Not reported

12

TABLE 1. Comparison of Web-user studies.

TABLE 2. Comparison of three traditional IR-user studies.

TABLE 3. Comparison of three OPAC-user studies.

Category	Wallace study	Peters study	Millsap & Ferl study
Number of Searches	4,134 searches	13,258 searches	1,045 sessions
Session length (number of queries per user per session)	Not reported	Not reported	One or less: 32.8% (343) Two–five: 43.8% (458) More than five: 23.4% (245)
Query Length (number of terms per query)	Two or less Terms: 75% (3,101) More than two terms: 25% (1,034)	Not reported	Not reported
Number of relevant documents viewed per session	Less than 25: 82.1% (3,394) More than 25: 17.9% (740)	Not reported	1 to 50: 80.7% (843) More than 50: 19.3% (202)
Number of queries by keyword	53.1% (2,197)	31.9% (4,229)	23.9% (250) of sessions contained one or more queries of this type
Number of queries by title	24.2% (1,000)	34.2% (4,534)	62.2% (650) of sessions contained one or more queries of this type
Number of queries by author	21.7% (897)	23.2% (3,076)	38.1% (398) of sessions contained one or more queries of this type
Use of advanced features (number of queries containing advanced options)	8.7% (360)	2.8% (371)	Not reported
Use of Boolean (number of queries containing Boolean operators)	Not reported	1% (133)	9.2% (96) of the sessions contained one or more queries of this type
Failure rate (number of queries improperly formatted)	7% (289)	15.3% (2,028)	10% (105) of the sessions contained one or more improperly formatted query

*Numbers not reported and not calculable based on data provided.

12

TABLE 1. Comparison of Web-user studies.

TABLE 2. Comparison of three traditional IR-user studies.

TABLE 4. Comparison of typical searches across three categories.

Category	Web systems searches	Traditional IR systems searches	OPAC systems searches
Session length (number of queries per user per session)	1–2	7–16	2–5
Query length (number of terms per query)	2	6–9	1–2
Number of relevant documents viewed per session	10 or less	Approximately 10	Less than 50
Use of advanced features (number of queries containing advanced options)	9%	9%	8%
Use of Boolean (number of queries containing Boolean operators)	8%	37%	1%
Failure rate (number of queries improperly formatted)	10%	17%	7–19%

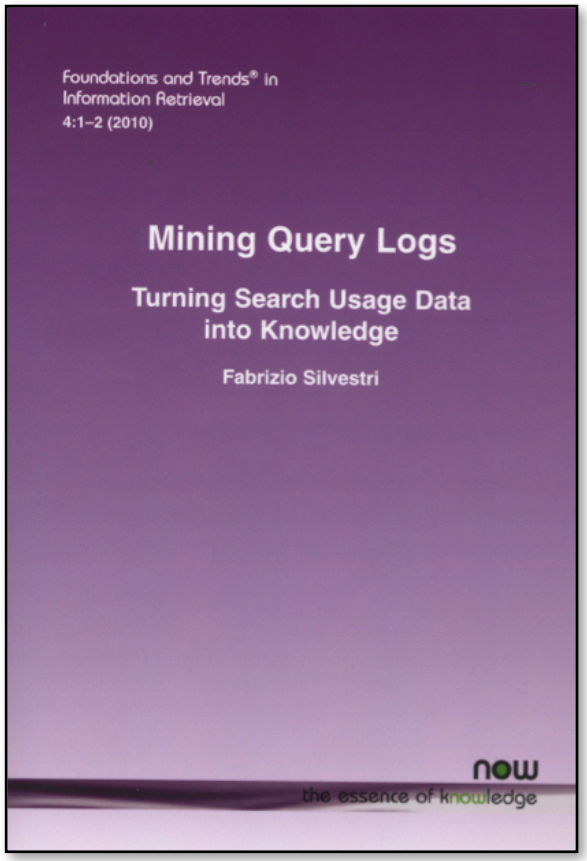
Failure rate (number of queries improperly formatted)	7% (289)	15.3% (2,028)	10% (105) of the sessions contained one or more improperly formatted query

*Numbers not reported and not calculable based on data provided.

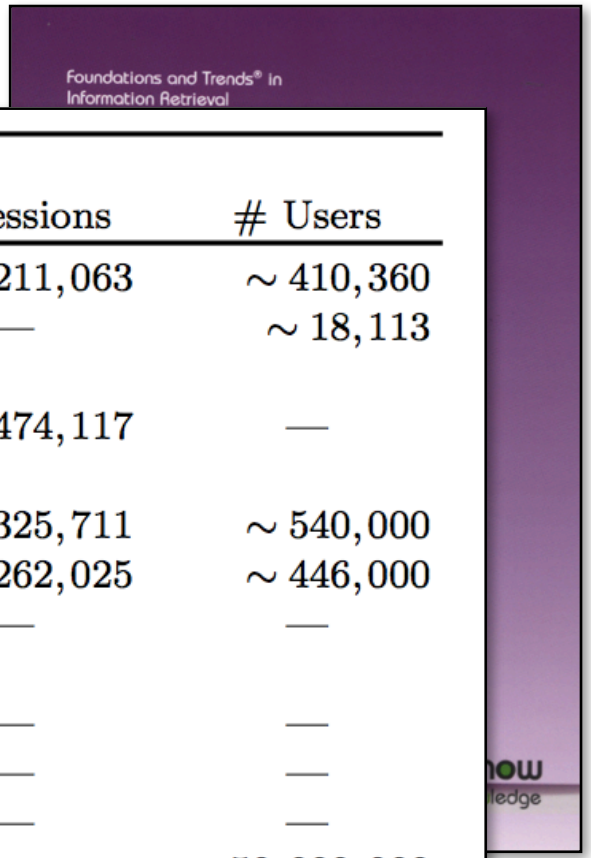
Jansen and Pooch, 2001

- Survey and comparison of log studies
  - Early web log studies
    - Fireball (German Web search engine, late 1990s)
    - Excite (Web serach engine, late 1990s)
    - Altavista (Web search engine, late 1990s)
  - Traditional IR systems
    - Hsieh-Yee (1993), Koenemann and Belkin (1996), Siegfried, Bates and Wilde (1993).
  - OPAC type systems
    - Millsap and Ferl (1993), Peters (1989), Wallace (1993)
- Main insights
  - Differences in manner Web users search versus searching characteristics of users on traditional IR or OPAC systems
  - Noticeable variation and use of metrics among studies

Silvestri (2010)

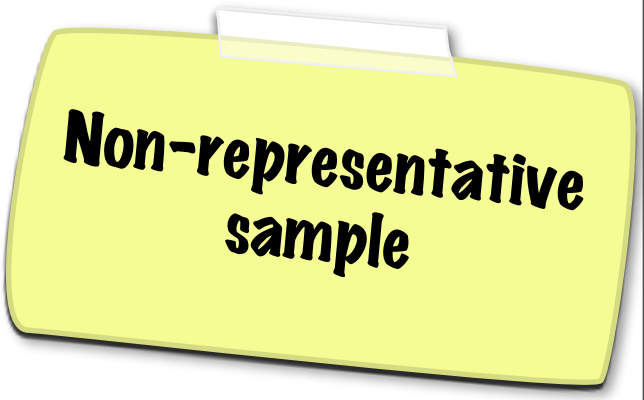


- Extensive survey, highly recommended
- Broad range over areas covered
  - Basic statistics
  - User interaction (sessions)
  - Efectiveness
  - Efficiency
  - New directions



Query log name	Public	Period	# Queries	# Sessions	# Users
Excite 1997	Y	Sep 1997	1,025,908	211,063	~ 410,360
Excite 1997 (small)	Y	Sep 1997	51,473	—	~ 18,113
Altavista	N	2 Aug–13 Sep 1998	993,208,159	285,474,117	—
Excite 1999	Y	Dec 1999	1,025,910	325,711	~ 540,000
Excite 2001	Y	May 2001	1,025,910	262,025	~ 446,000
Altavista (public)	Y	Sep 2001	7,175,648	—	—
Tiscali	N	Apr 2002	3,278,211	—	—
TodoBR	Y	Jan–Oct 2003	22,589,568	—	—
TodoCL	N	May–Nov 2003	—	—	—
AOL (big)	N	Dec 26 2003–Jan 1 2004	~ 100,000,000	—	~ 50,000,000
Yahoo!	N	Nov 2005–Nov 2006	—	—	—
AOL (small)	Y	1 Mar–31 May 2006	36,389,567	—	—

Some specialized log file analysis



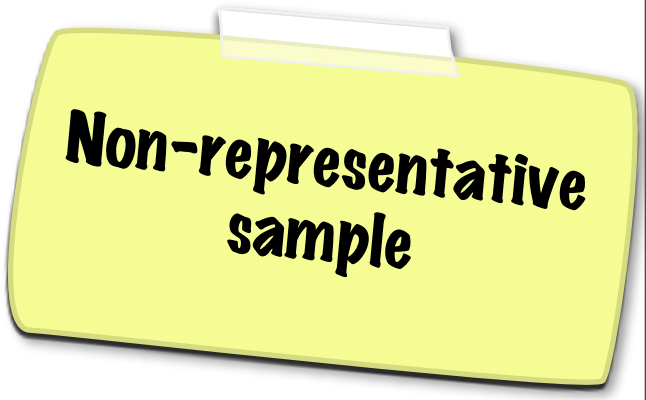
Non-representative sample

Some specialized log file analysis

- Mishne and de Rijke (2006)
  - Study the behavior of users of a blog search engine through a log file analysis
  - Different queries from web users? Different sessions?

14

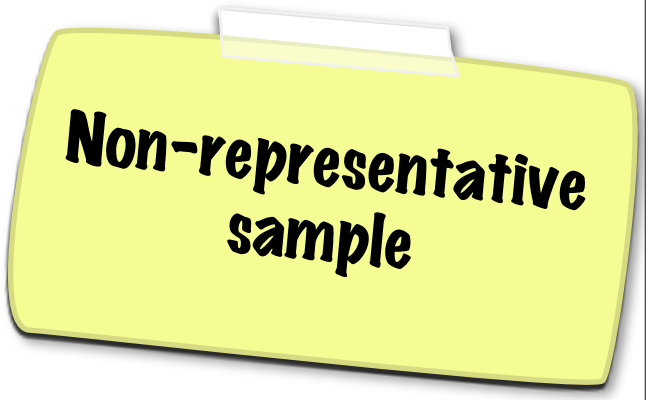
Some specialized log file analysis



- Mishne and de Rijke (2006)
  - Study the behavior of users of a blog search engine through a log file analysis
  - Different queries from web users? Different sessions?
- Carman et al. (2009)
  - Examine difference between vocabularies of queries, social bookmarking tags, and online documents
    - Look at different domains, correlations between queries and tags

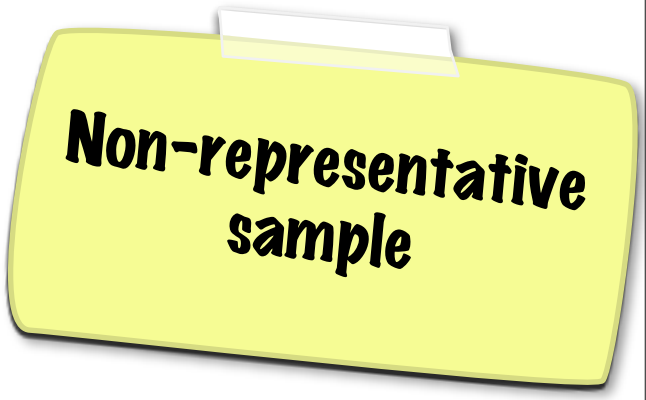
14

Some specialized log file analysis



- Mishne and de Rijke (2006)
  - Study the behavior of users of a blog search engine through a log file analysis
  - Different queries from web users? Different sessions?
- Carman et al. (2009)
  - Examine difference between vocabularies of queries, social bookmarking tags, and online documents
    - Look at different domains, correlations between queries and tags
- Huurnink et al. (2010)
  - Search behavior of media professionals
  - Map queries to categories associated to documents

Some specialized log file analysis



- Mishne and de Rijke (2006)
  - Study the behavior of users of a blog search engine through a log file analysis
  - Different queries from web users? Different sessions?
- Carman et al. (2009)
  - Examine difference between vocabularies of queries, social bookmarking tags, and online documents
    - Look at different domains, correlations between queries and tags
- Huurnink et al. (2010)
  - Search behavior of media professionals
  - Map queries to categories associated to documents
- Weerkamp et al. (2011)
  - Search behavior at a people search engine
  - Meta-search, very high degrees of ambiguity

A closer look at people search

- Weerkamp et al., People Searching for People: Analysis of a People Search Engine Log. *SIGIR 2011*.

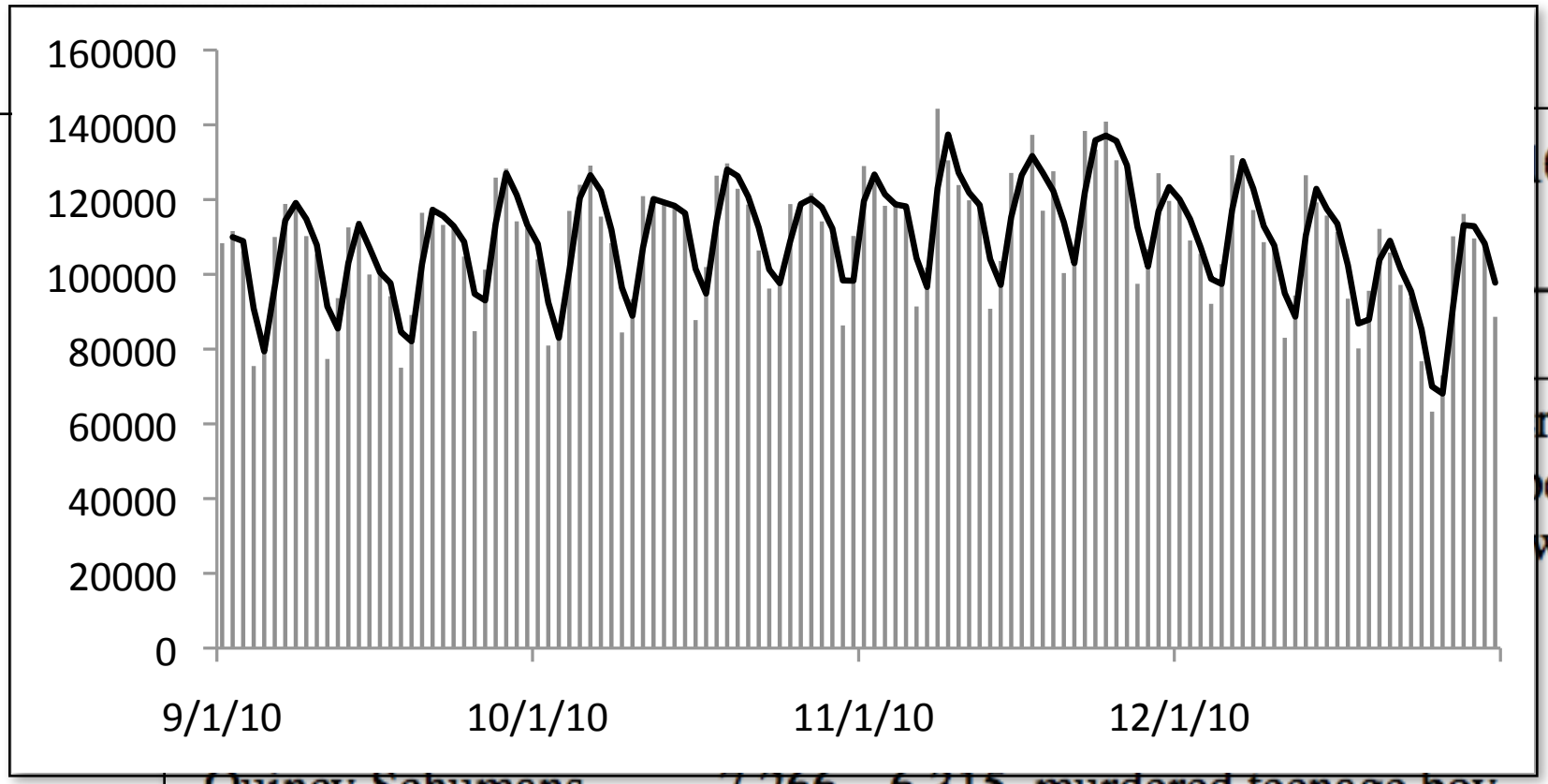
15

Table 2: Characteristics of individual queries.

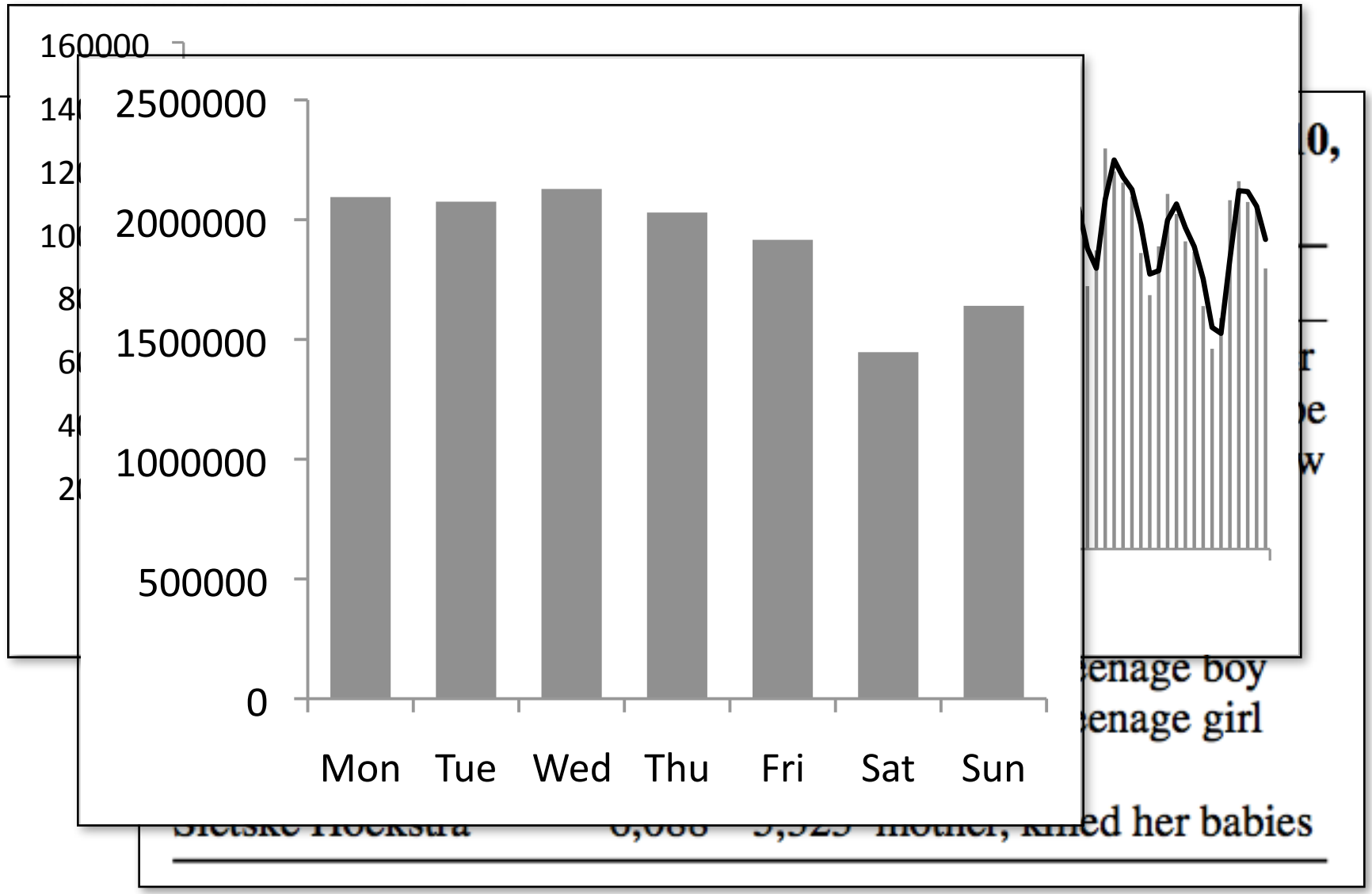
Number of queries	13,331,417	
Number of unique queries	4,221,556	
Number of single-term queries	537,365	(4.0%)
Average number of queries per day	110,177	
Busiest day in number of queries	144,309	
Number of queries with keyword	514,850	(3.9%)

Table 3: 10 most popular queries during Sep. 1–Dec. 31, 2010, in terms of query counts and unique users.

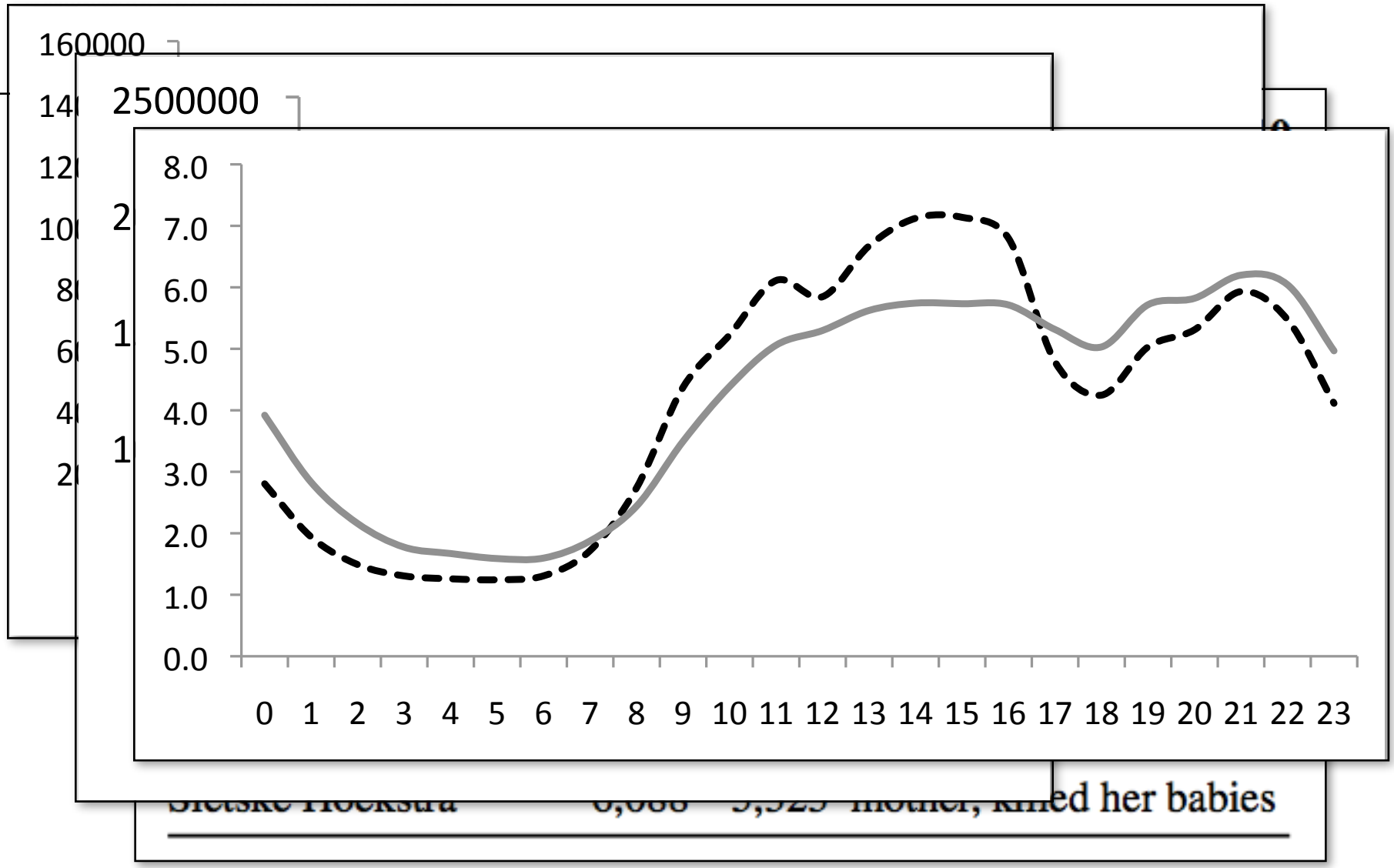
Name	Count	Users	Gloss
Suze van Rozelaar	16,929	15,373	mistress of soccer player
Kelly Huizen	13,005	11,706	teenage girl with sex tape
Ben Saunders	10,074	9,145	participant of talent show
Barbara van der Vegte	9,879	8,256	mistress of tv host
Geert Wilders	8,990	8,483	politician
Lieke van Lexmond	7,774	6,368	actress
Quincy Schumans	7,266	6,315	murdered teenage boy
Joyce Exalto	6,656	5,584	murdered teenage girl
Aa Aa	6,457	6,442	test query
Sietske Hoekstra	6,088	5,323	mother, killed her babies



Quincy Schumans	7,266	6,315	murdered teenage boy
Joyce Exalto	6,656	5,584	murdered teenage girl
Aa Aa	6,457	6,442	test query
Sietske Hoekstra	6,088	5,323	mother, killed her babies


160000
2500000
2000000
1500000
1000000
500000
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
enage boy
enage girl
ed her babies

15


160000
2500000
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
ed her babies

15

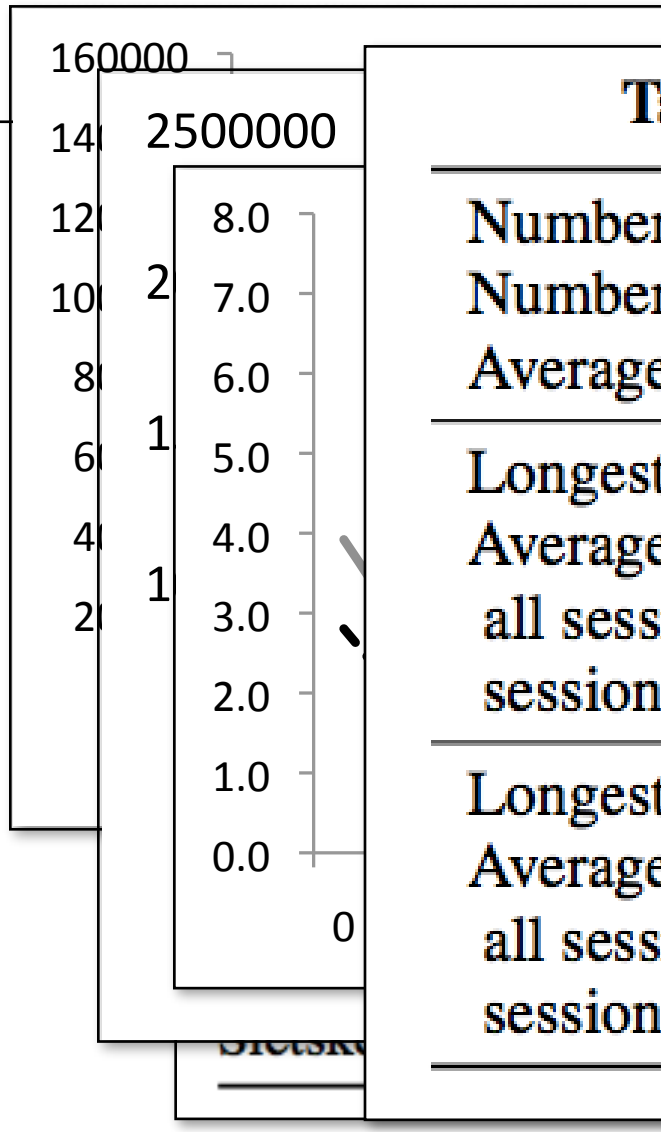


Table 4: Characteristics of sessions.

Number of sessions	8,125,695
Number of sessions with > 1 query	1,775,880
Average number of sessions per day	67,155
Longest session in hours	08h25m
Average session duration	
all sessions	1m21s
sessions with > 1 query	6m9s
Longest session in number of queries	1,302
Average session length	
all sessions	1.64
sessions with > 1 query	3.93

15

160000
14
12
10
8
6
4
2

Table 5: Characteristics of users.

Number of users	6,841,442
Number of users with > 1 query	1,481,377
Number of users with > 1 session	514,042
Busiest day in unique users	11/24/2010 90,799
Average number of queries per user	
all users	1.95
users with > 1 query	5.38
Average number of sessions per user	
all users	1.19
users with > 1 session	3.50

15

160000

14

Table 4: Characteristics of sessions

Table 5: Characteristics of users

Table 6: Characteristics of out clicks.

Number of out clicks	3,965,462	
Number of unique out clicks	2,883,230	
Number of queries followed by out click	2,351,848	17.6%
Number of sessions that include out click	1,625,817	20.0%

users with > 1 query 3.38

Average number of sessions per user	
all users	1.19
users with > 1 session	3.50

160000

14

Table 4: Characteristics of sessions

Table 5: Characteristics of users

Table 6: Characteristics of out clicks.

Table 7: Interface result categories and number of out clicks.

Social media	2,625,500	66.2%
Search engines	674,079	17.0%
Multimedia	120,874	3.1%
Miscellaneous	337,104	8.5%
"Alternative sources"	187,098	4.7%

all users	1.19
users with > 1 session	3.50

15

160000

14

Table 4: Characteristics of sessions

Table 5: Characteristics of users

Table 6: Characteristics of out clicks

Table 7: Interface result categories and number of out clicks

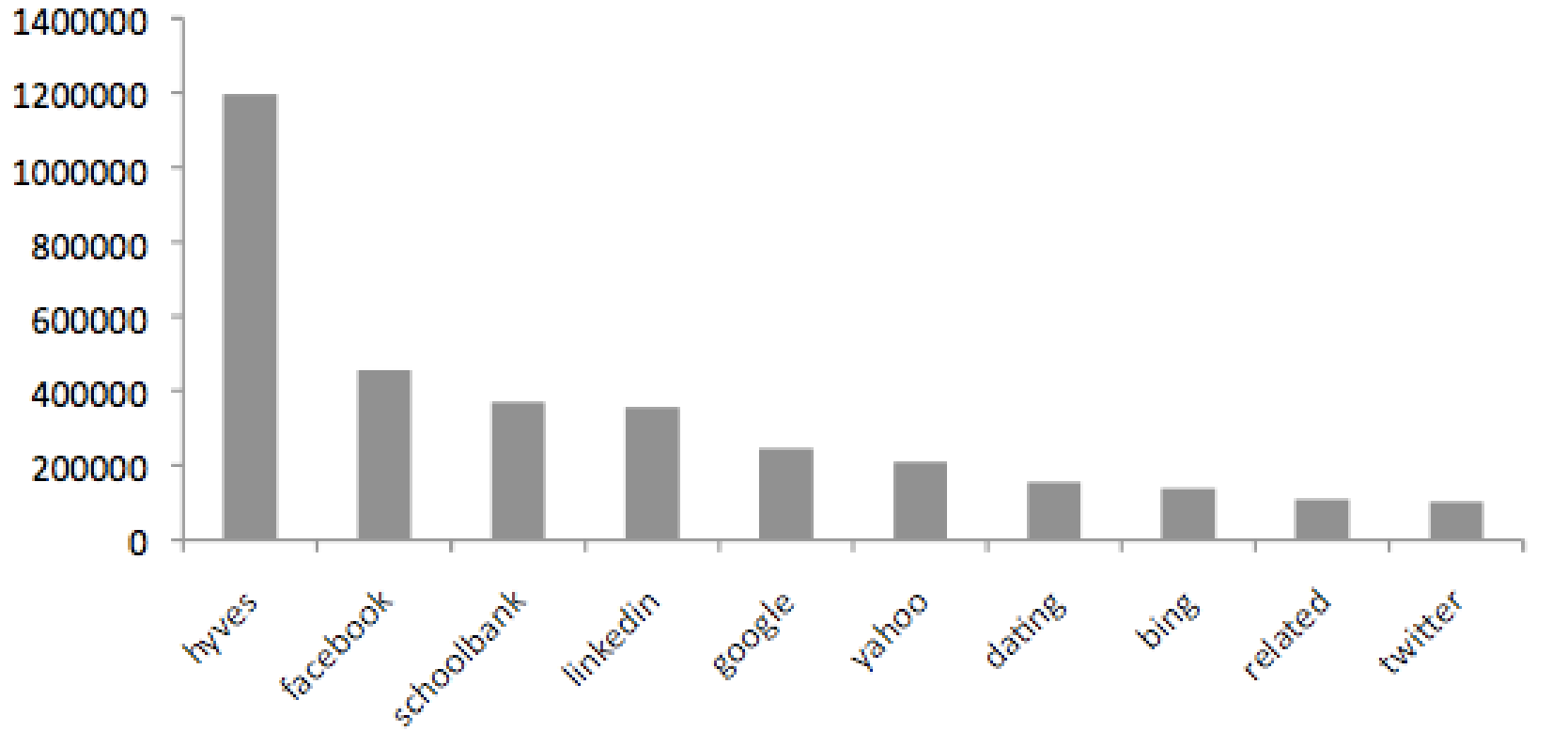
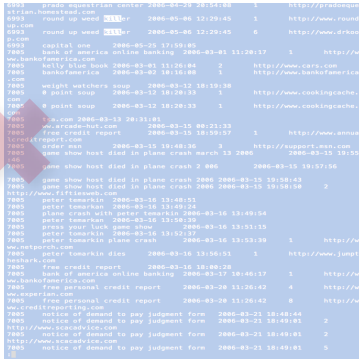


Figure 7: Number of out clicks per result type.

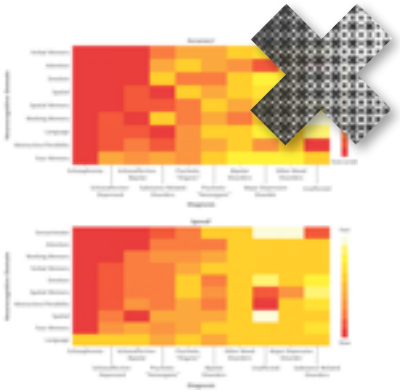
15

UNIVERSITEIT VAN AMSTERDAM

What



Examples



Analysis

Uses

What to do?

Wrap-up

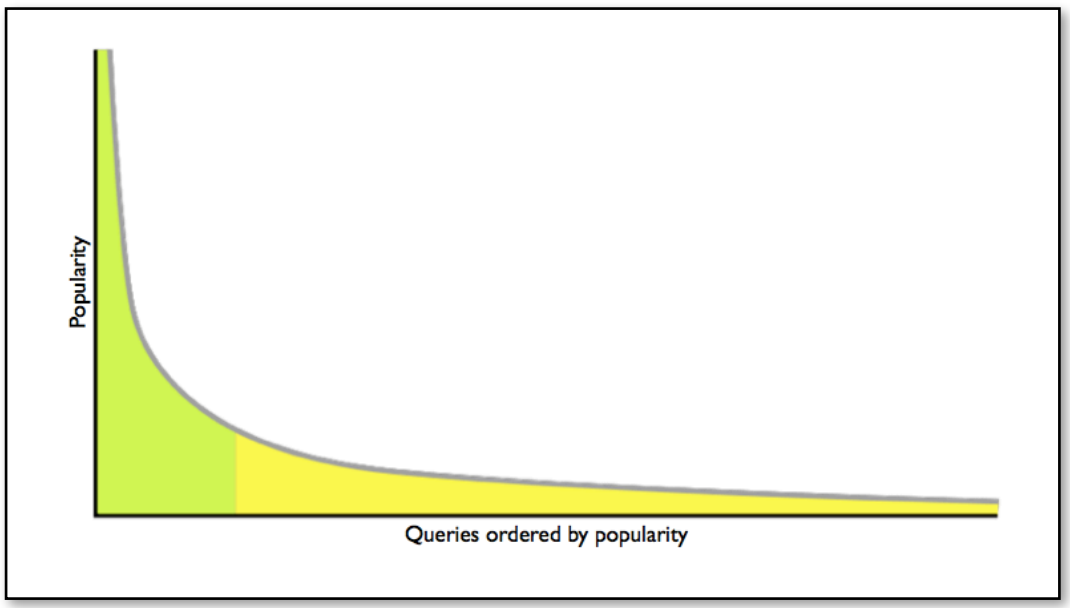
Making sense

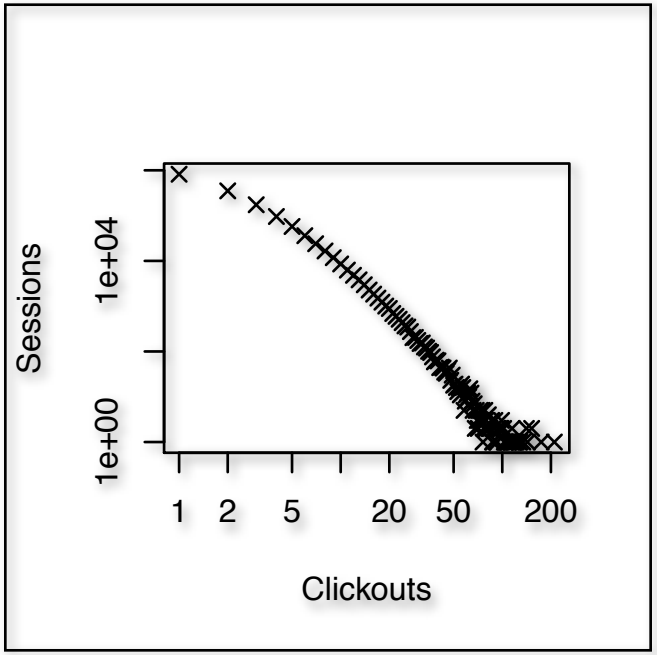
- Patterns
- Regularities
- Categories
- Intent
- Longitudinal aspects
- No session is an island

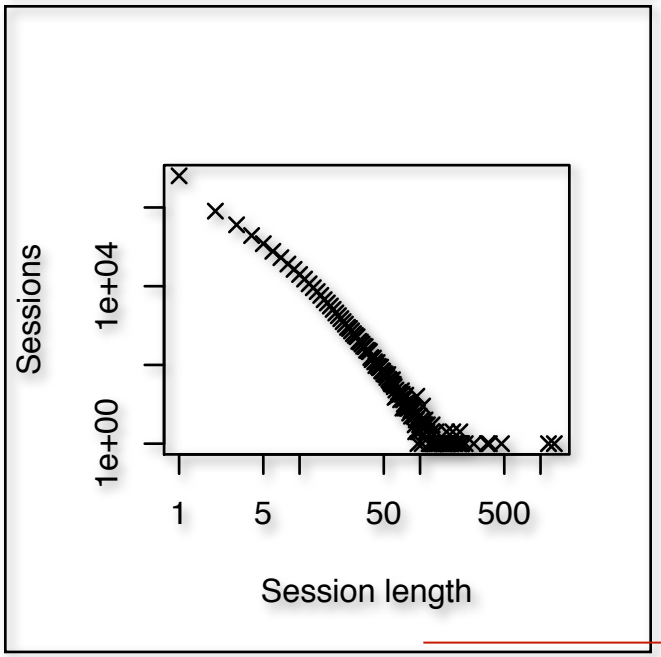
UNIVERSITEIT VAN AMSTERDAM

Power laws

- Power laws can be found everywhere in log files







Categories

- What are those queries about?
  - Mishne and De Rijke use Froogle and Yahoo product search
    - Yahoo! directory
      - Use category of top page retrieved: Yahoo! category
    - Froogle
      - Use top shopping category: Froogle category
  - “Map queries to pages for which you have category information, and let these pages vote”
  - Coverage: 55% (Yahoo!), 68% (Froogle)

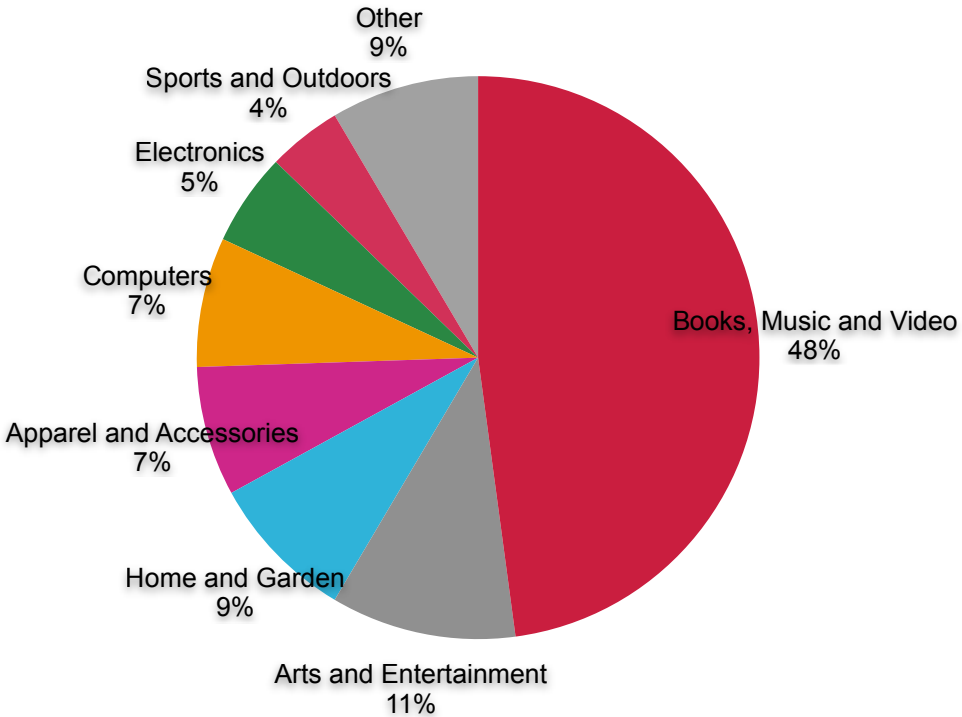
19

UNIVERSITEIT VAN AMSTERDAM

Categories

Query: **24** *Yahoo! category*: /Entertainment/Television Shows/Action and Adventure/24 *Froogle category*: /Books, Music and Video/Video/Action and Adventure
Query: **Atkins** *Yahoo! category*: /Business and Economy/Shopping and Services/Health/Weight Loss/Diets and Programs/Low Carbohydrate Diets/Atkins Nutritional Approach *Froogle category*: /Food and Gourmet/Food/Snack Foods
Query: **Evolution debate** *Yahoo! category*: /Society and Culture/Religion and Spirituality/Science and Religion/Creation vs. Evolution/Intelligent Design *Froogle category*: /Books, Music and Video/Books/Social Sciences
Query: **Vioxx** *Yahoo! category*: /Health/Pharmacy/Drugs and Medications/Specific Drugs and Medications/Vioxx, Rofecoxib *Froogle category*: /Health and Personal Care/Over-the-Counter Medicine

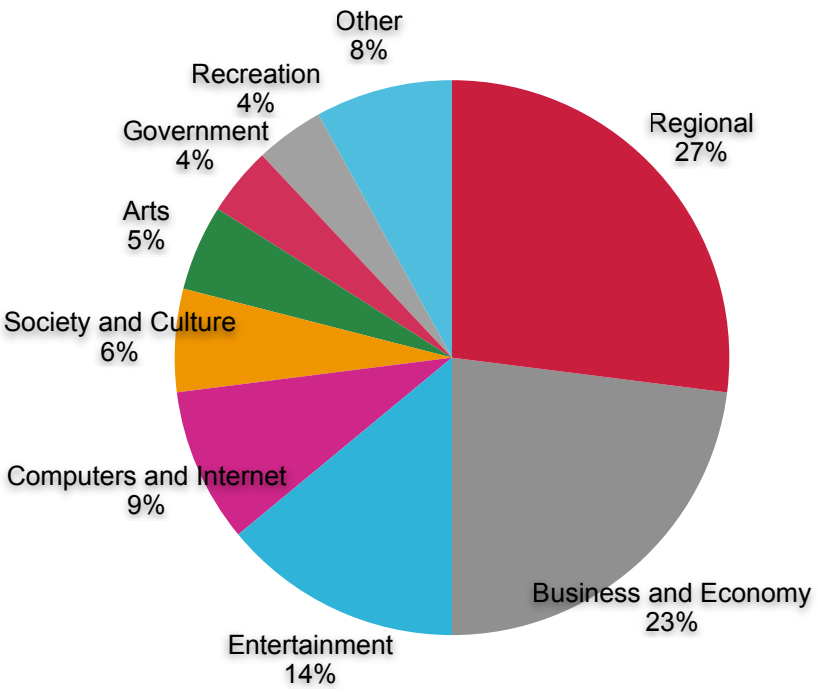
Ad-hoc — Top Yahoo! (L) and Froogle (R)

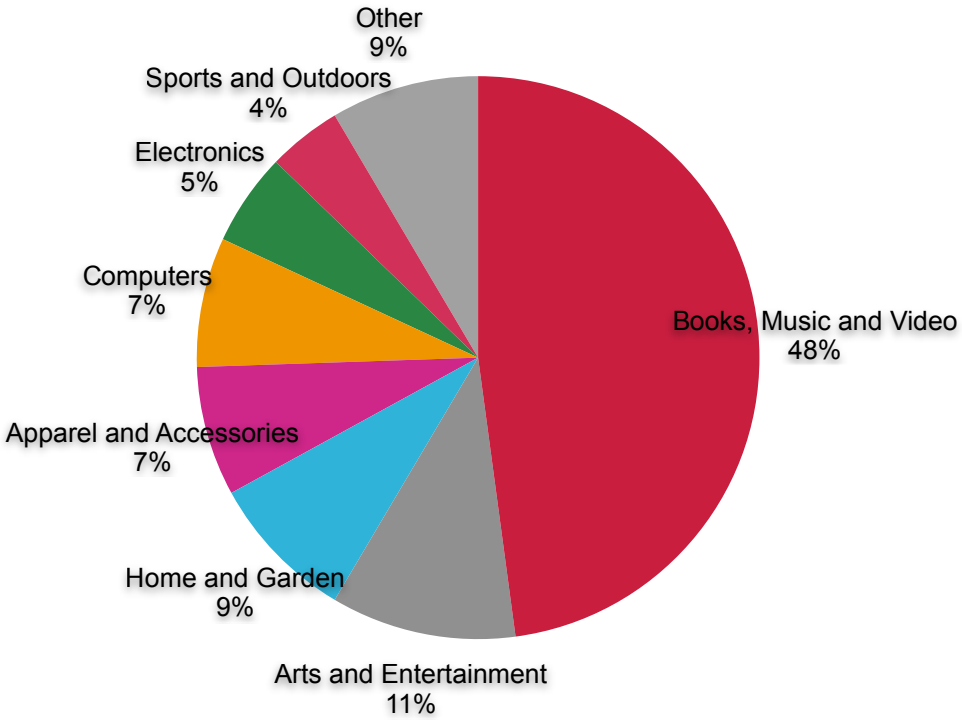


- Sample findings
  - In terms of interest areas, blog searchers are more engaged in technology and politics than web searchers (2005!)
  - Noticeable interest in named entities: people, brands, companies, etc.

UNIVERSITEIT VAN AMSTERDAM

Ad-hoc — Top Yahoo! (L) and Froogle (R)





- Sample findings
  - In terms of interest areas, blog searchers are more engaged in technology and politics than web searchers (2005!)
  - Noticeable interest in named entities: people, brands, companies, etc.

UNIVERSITEIT VAN AMSTERDAM

Categories (2)

Categories (2)

- What are those queries about?
  - Meij et al. 2009
  - Setting: *Sound and Vision*
    - Media professionals
  - Assign title of Wikipedia pages
    - “A page is a concept”

UNIVERSITEIT VAN AMSTERDAM

Categories (2)

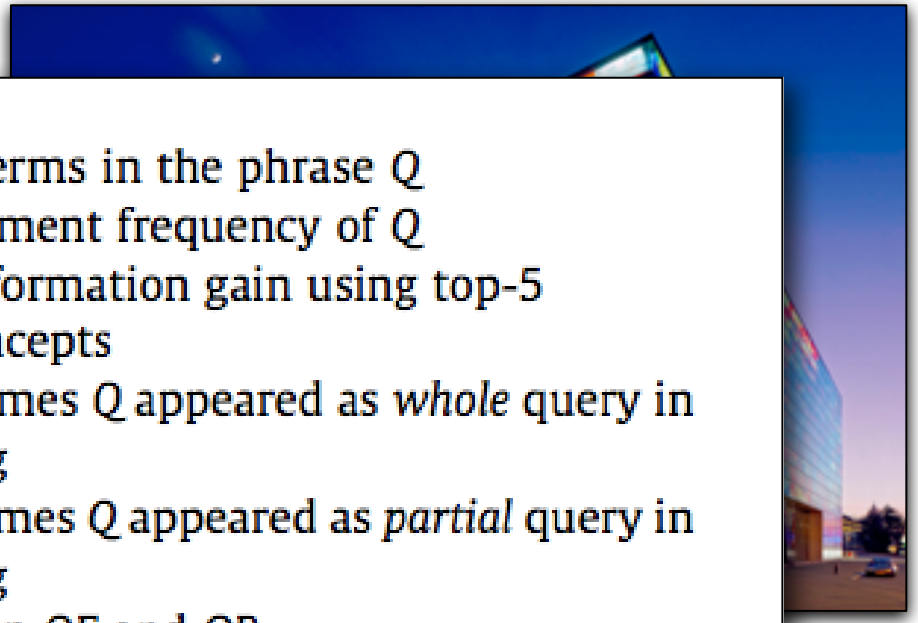
- What are those queries about?
  - Meij et al. 2009
  - Setting: *Sound and Vision*
    - Media professionals
  - Assign title of Wikipedia pages
    - "A page is a concept"
  - Approach retrieval + learning
    - The tasks becomes: assign a concept from a background knowledge source to an incoming query

Categories (2)

- What are those queries about?
  - Meij et al. 2009
  - Setting: *Sound and Vision*
    - Media professionals
  - Assign title of Wikipedia pages
    - "A page is a concept"
  - Approach retrieval + learning
    - The tasks becomes: assign a concept from a background knowledge source to an incoming query
  - Retrieval
    - Given an incoming query for the archive, assign concept
    - Query strings + session
  - Learning
    - Take the top 5 produced by the retrieval step, learn to remove non-relevant concepts

21

UNIVERSITEIT VAN AMSTERDAM



N-gram features

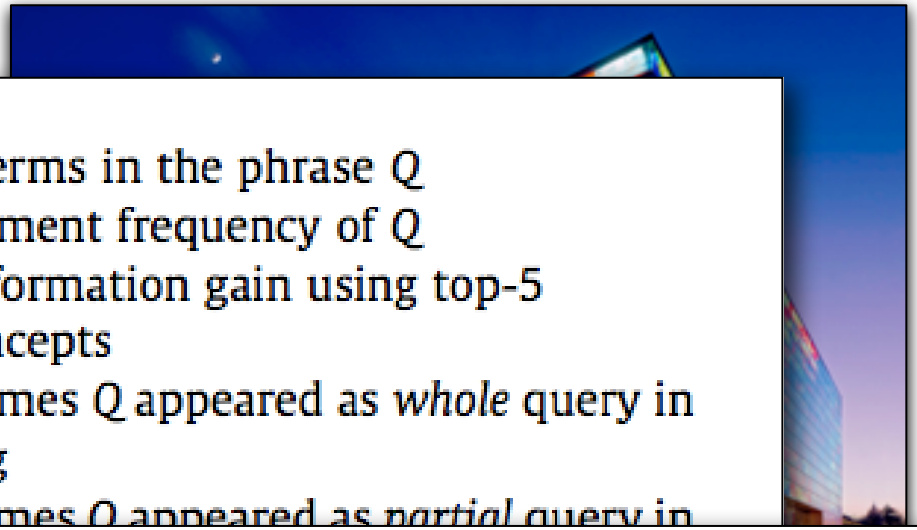
$LEN(Q) = \|Q\|$	Number of terms in the phrase Q
$IDF(Q)$	Inverse document frequency of Q
$WIG(Q)$	Weighted information gain using top-5 retrieved concepts
$QE(Q)$	Number of times Q appeared as *whole* query in the query log
$QP(Q)$	Number of times Q appeared as *partial* query in the query log
$QEQP(Q)$	Ratio between QE and QP
$SNIL(Q)$	Does a sub-n-gram of Q fully match with any concept label?
$SNCL(Q)$	Is a sub-n-gram of Q contained in any concept label?

- Given an incoming query for the archive, assign concept
- Query strings + session

- Learning
  - Take the top 5 produced by the retrieval step, learn to remove non-relevant concepts

21

UNIVERSITEIT VAN AMSTERDAM



N-gram features

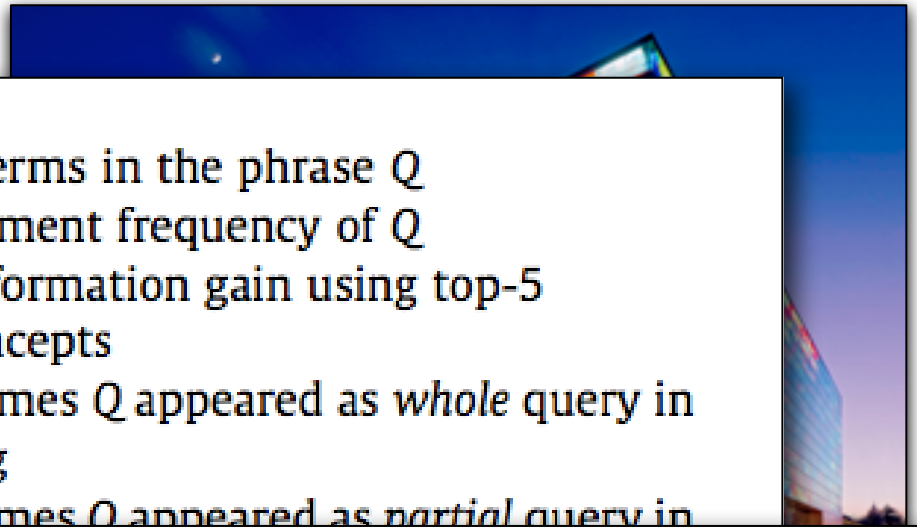
$LEN(Q) = \|Q\|$	Number of terms in the phrase Q
$IDF(Q)$	Inverse document frequency of Q
$WIG(Q)$	Weighted information gain using top-5 retrieved concepts
$QE(Q)$	Number of times Q appeared as *whole* query in the query log
$QP(Q)$	Number of times Q appeared as *partial* query in

Concept features

$INLINKS(c)$	The number of concepts linking to c
$OUTLINKS(c)$	The number of concepts linking from c
$GEN(c)$	Function of depth of c in the SKOS category hierarchy
$CAT(c)$	Number of associated categories
$REDIRECT(c)$	Number of redirect pages linking to c

- Learning
  - Take the top 5 produced by the retrieval step, learn to remove non-relevant concepts

UNIVERSITEIT VAN AMSTERDAM



N-gram features

$LEN(Q) = \|Q\|$	Number of terms in the phrase Q
$IDF(Q)$	Inverse document frequency of Q
$WIG(Q)$	Weighted information gain using top-5 retrieved concepts
$QE(Q)$	Number of times Q appeared as *whole* query in the query log
$QP(Q)$	Number of times Q appeared as *partial* query in

Concept featu

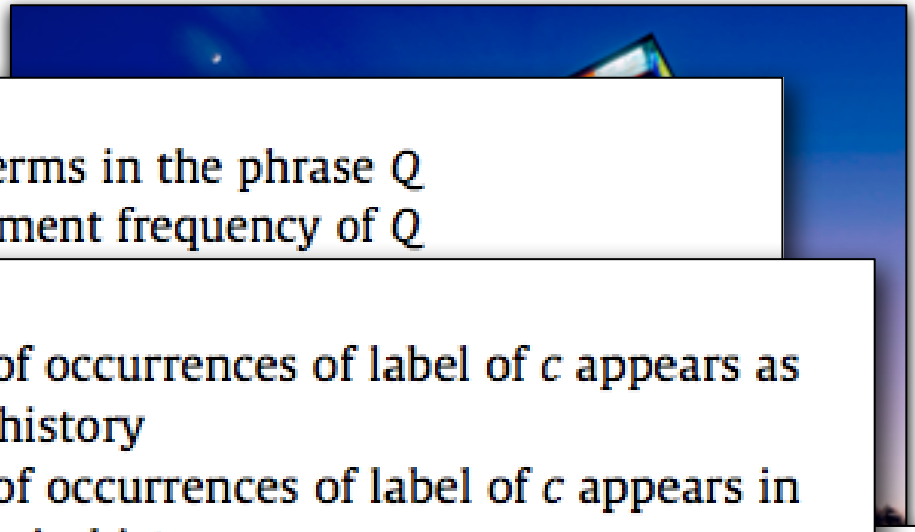
$INLINKS(c)$	
$OUTLINKS(c)$	
$GEN(c)$	
$CAT(c)$	
$REDIRECT(c)$	

N-gram + concept features

$TF(c,Q) = \frac{n(Q,c)}{\|c\|}$	Relative phrase frequency of Q in c by length of c
$TF_f(c,Q) = \frac{n(Q,c,f)}{\|f\|}$	Relative phrase frequency of Q in representationfof c normalized by
$POS_n(c,Q) = pos_n(Q)/\|c\|$	Position of nth occurrence of Q in c by length of c
$SPR(c,Q)$	Spread (distance between the last occurrences of Q in c)
$TF \cdot IDF(c,Q)$	The importance of Q for c
$RIDF(c,Q)$	Residual IDF (difference between e observed IDF)

- Learning
  - Take th relevan

UNIVERSITEIT VAN AMSTERDAM



N-gram features

$LEN(Q) = \|Q\|$	Number of terms in the phrase Q
$IDF(Q)$	Inverse document frequency of Q

History features

$CCIH(c)$	Number of occurrences of label of c appears as query in history
$CCCH(c)$	Number of occurrences of label of c appears in any query in history
$CIHH(c)$	Number of times c is retrieved as result for any query in history
$CCIHH(c)$	Number of times label of c equals title of any result for any query in history
$CCCHH(c)$	Number of times title of any result for any query in history contains label of c
$QCIHH(Q)$	Number of times title of any result for any query in history equals Q
$QCCHH(Q)$	Number of times title of any result for any query in history contains Q
$QCIH(Q)$	Number of times Q appears as query in history
$QCCH(Q)$	Number of times Q appears in any query in history

UNIVERSITEIT VAN AMSTERDAM

N-gram features

$LEN(Q) = \|Q\|$	Number of terms in the phrase Q
$IDF(Q)$	Inverse document frequency of Q

History features

$CCIH(c)$	Number of occurrences of label of c appears as query in history
$CCCH(c)$	Number of occurrences of label of c appears in
CI	
CC	
CC	
QC	
QC	
	query in history contains Q
$QCIH(Q)$	Number of times Q appears as query in history
$QCCH(Q)$	Number of times Q appears in any query in history

Results for full query-based reranking. ▲, ▼ and ° indicate that a score is significantly better, worse or statistically indistinguishable, respectively. The leftmost symbol represents the difference with the baseline, the next with the J48 run, and the rightmost with the NB run.

	P1	R-prec	Recall	MRR	SR
Baseline	0.5636	0.5216	0.6768	0.6400	0.7535
J48	0.7152▲	5857°	0.6597°	0.6877°	0.7317°
NB	0.6925▲°	0.5897°°	0.6865°°	0.6989°°	0.7626°°
SVM	**0.8833▲▲▲**	**0.8666▲▲▲**	**0.8975▲▲▲**	**0.8406▲°▲**	**0.9053▲▲▲**

UNIVERSITEIT VAN AMSTERDAM

Intent

Intent

- “The need behind the query”

22

Intent

- "The need behind the query"
- Smeulders et al. Content-based image retrieval at the end of the early years, *IEEE TPAMI*, 2000
  - Content-based image retrieval queries
    - Target (or known-item) search (when the user has a specific image in mind)
    - Category search (retrieving an arbitrary image representative of a specific class)
    - Search by association (search starts with no aim other than to find interesting things)

Intent

- "The need behind the query"
- Smeulders et al. Content-based image retrieval at the end of the early years, *IEEE TPAMI*, 2000
  - Content-based image retrieval queries
    - Target (or known-item) search (when the user has a specific image in mind)
    - Category search (retrieving an arbitrary image representative of a specific class)
    - Search by association (search starts with no aim other than to find interesting things)
- Broder, Taxonomy of web search. *SIGIR Forum*, 2003
  - Informational (I need to know about a topic)
  - Navigational (Take me to a specific item or site)
  - Transactional (Download a product or service)

22

Intent

- "The need behind the
- Smeulders et al. Cont
  early years, *IEEE TPA*
  - Content-based im
    - Target (or known
      in mind)
    - Category search
      specific class)
    - Search by assoc
      interesting thing
- Broder, Taxonomy of web search. *SIGIR Forum*, 2003
  - Informational (I need to know about a topic)
  - Navigational (Take me to a specific item or site)
  - Transactional (Download a product or service)

Type of query	User Survey	Query Log Analysis
Navigational	24.5%	20%
Informational	?? (estimated 39%)	48%
Transactional	> 22% (estimated 36%)	30%

Figure 5. Query classification.

Intent (2)

- Broder's taxonomy used, refined and modfied by many
- Rose and Levinson (2004)
  - ~~Transactional~~ → Resource
  - Refinement of Informational and Resource

UNIVE

Inte

- B
- R

1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (I.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

23

UNIVE

1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book

Table 3: Results of Classifying Queries by Search Goals

GOAL		SET 1	SET 2	SET 3
	directed	2.70%	3.30%	7.30%
	undirected	31.30%	26.50%	22.70%
	advice	2.00%	2.70%	5.00%
	locate	24.30%	25.90%	24.40%
	list	2.70%	2.90%	2.10%
informational total		**63.00%**	**61.30%**	**61.50%**
	download	4.30%	4.30%	5.60%
	entertain	4.00%	8.20%	5.80%
	interact	5.70%	4.30%	6.00%
	obtain	7.70%	10.30%	7.70%
resource total		**21.70%**	**27.00%**	**25.00%**
navigational		**15.30%**	**11.70%**	**13.50%**

at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.

house document no. 587

23

UNIVE

1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book

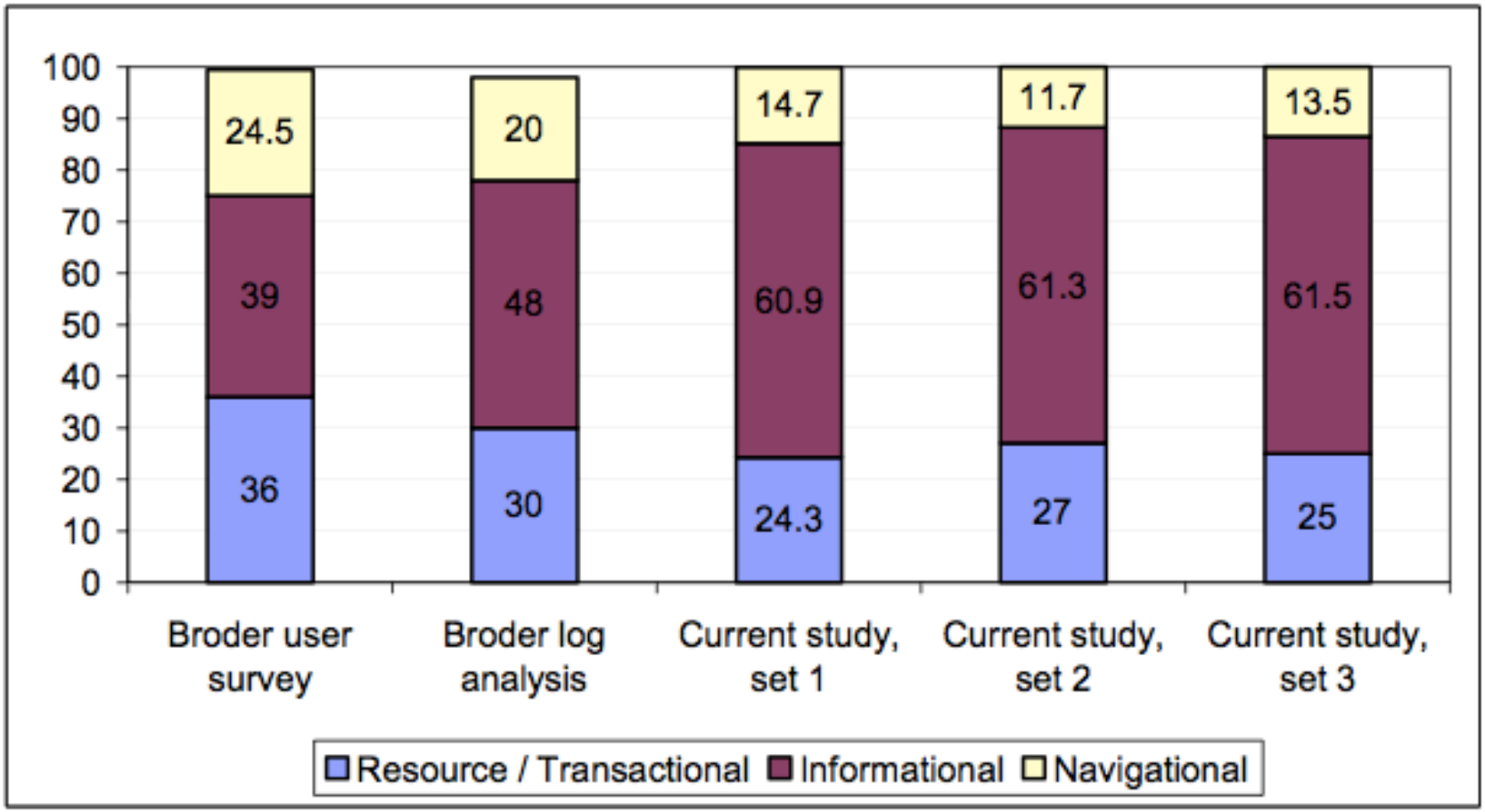


Figure 2: Comparison of Broder's search taxonomy to our top-level goals. Resource and informational results in the first column are Broder's estimates. Results do not total 100% due to rounding error.

UNIVERSITEIT VAN AMSTERDAM

Intent (3)

Intent (3)

- Variations

24

Intent (3)

- Variations
  - Blog search
    - **Context** queries: locate contexts in which a name appears in the blogspace (“iPhone 4S”)
    - **Concept** queries: locate blogs or blog posts that deal the searcher’s interests (“Euro crisis”)

24

Intent (3)

- Variations
  - Blog search
    - **Context** queries: locate contexts in which a name appears in the blogspace (“iPhone 4S”)
    - **Concept** queries: locate blogs or blog posts that deal the searcher’s interests (“Euro crisis”)
  - People search
    - High profile
      - Event-based ~ tracking
      - Regular ~ discovery
    - Low profile ~ catching up, discovery

24

Intent (3)

Table 9: Subclasses of the event-based high-profile queries and their percentage.

Event-based subclass	Percentage
Deaths	33.3%
Criminals	22.9%
Related to celebrities	9.7%
Related to other high-profiles	9.7%
Television	9.0%
Sex related	6.3%
Miscellaneous	9.0%

Intent (3)

- Variations
  - Blog search
    - **Context** queries: locate contexts in which a name appears in the blogspace (“iPhone 4S”)
    - **Concept** queries: locate blogs or blog posts that deal the searcher’s interests (“Euro crisis”)
  - People search
    - High profile
      - Event-based ~ tracking
      - Regular ~ discovery
    - Low profile ~ catching up, discovery
    - The need behind sessions
      - Family session
      - CV session
      - Event session
      - Spotting session
      - …

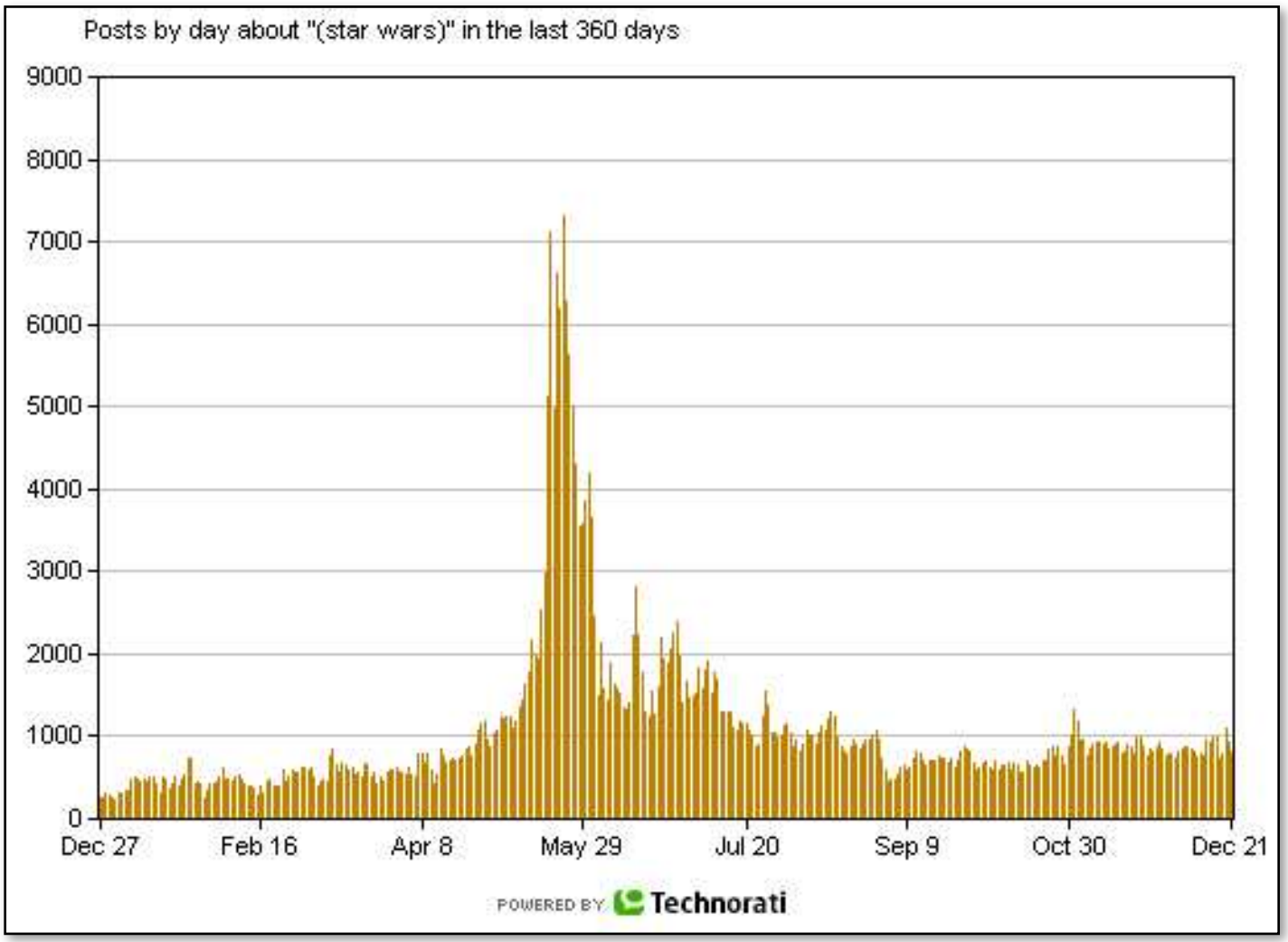
24

UNIVERSITEIT VAN AMSTERDAM

No session is an island

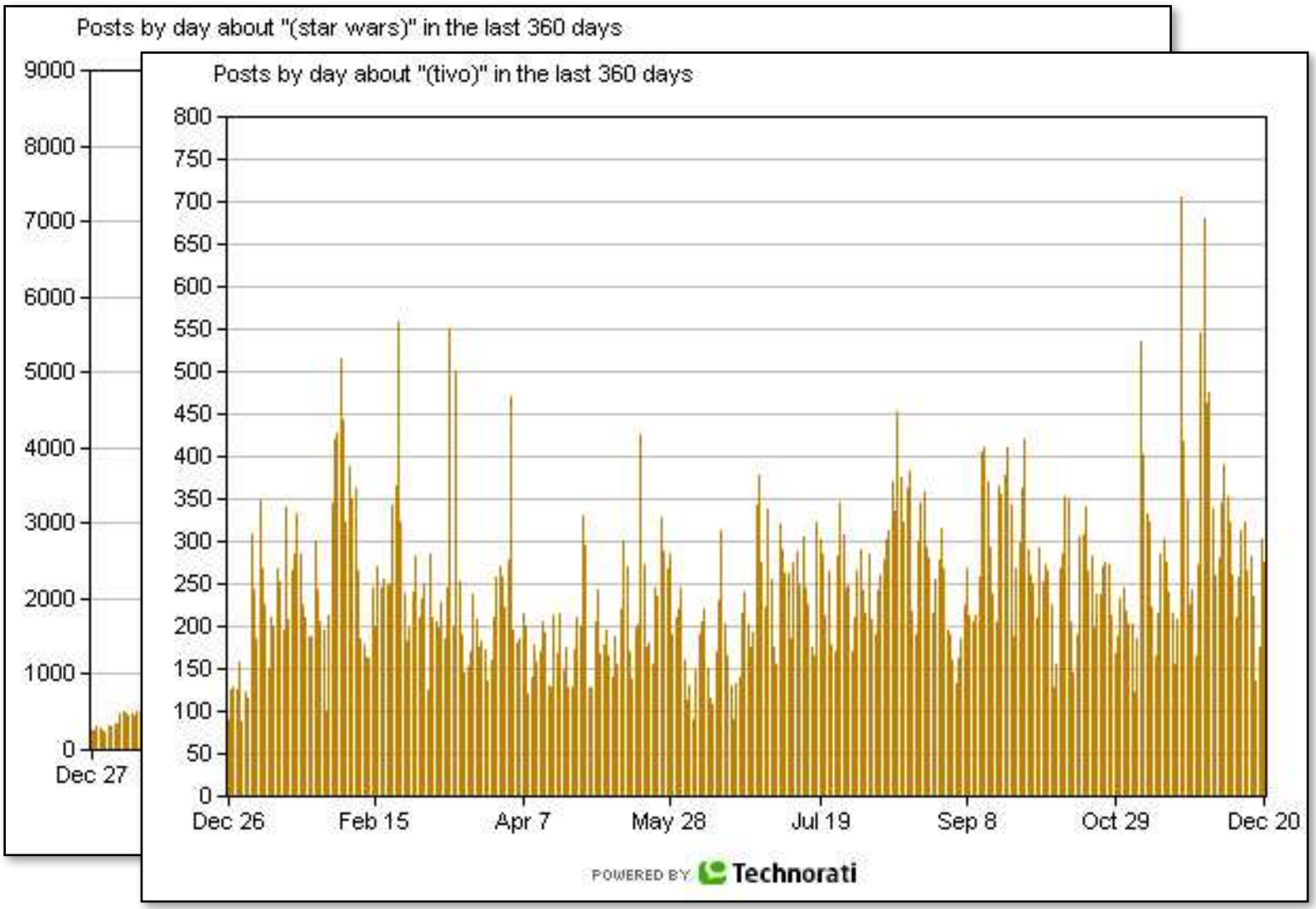
25

UNIVERSITEIT VAN AMSTERDAM



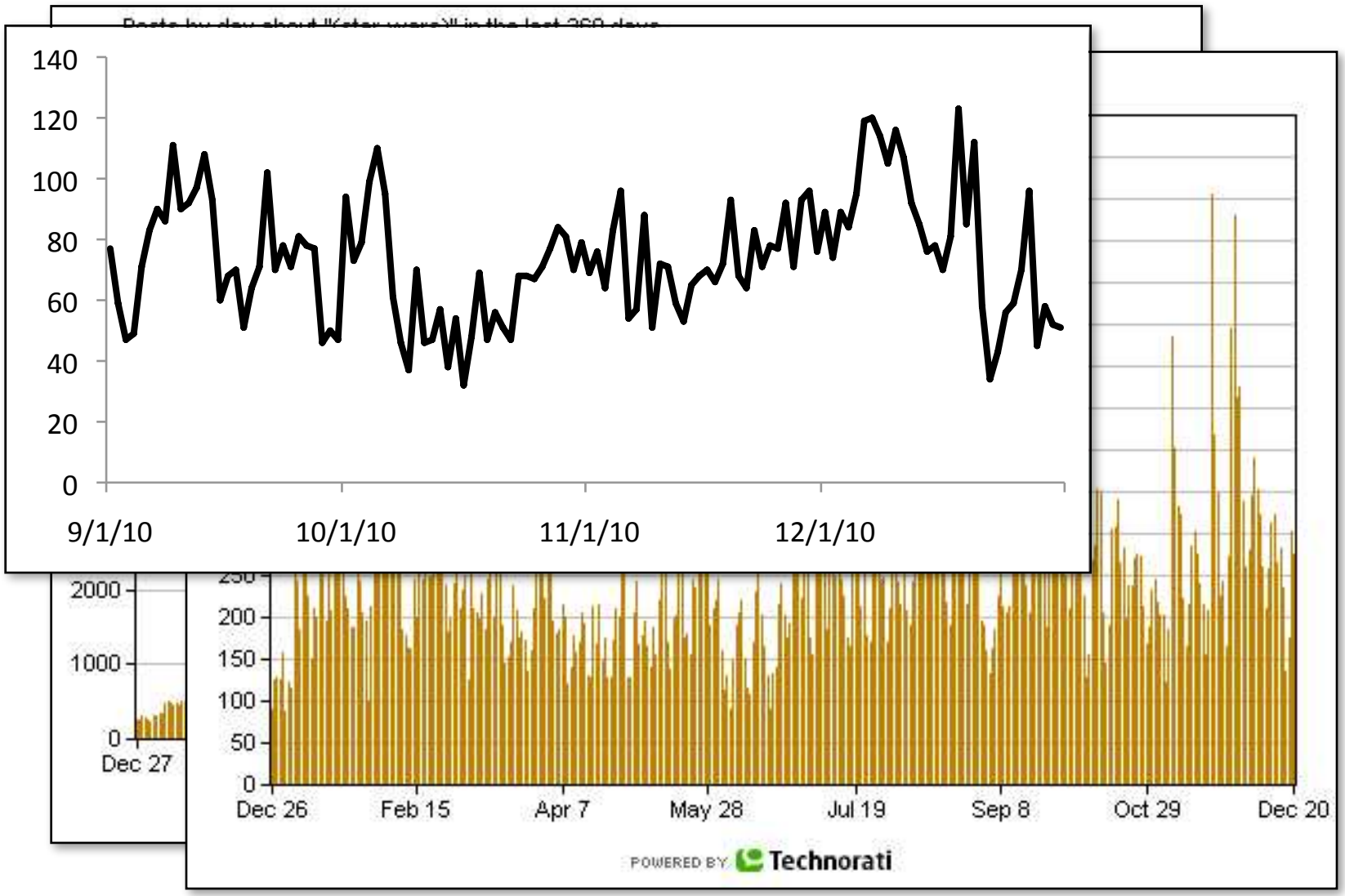
25

UNIVERSITEIT VAN AMSTERDAM



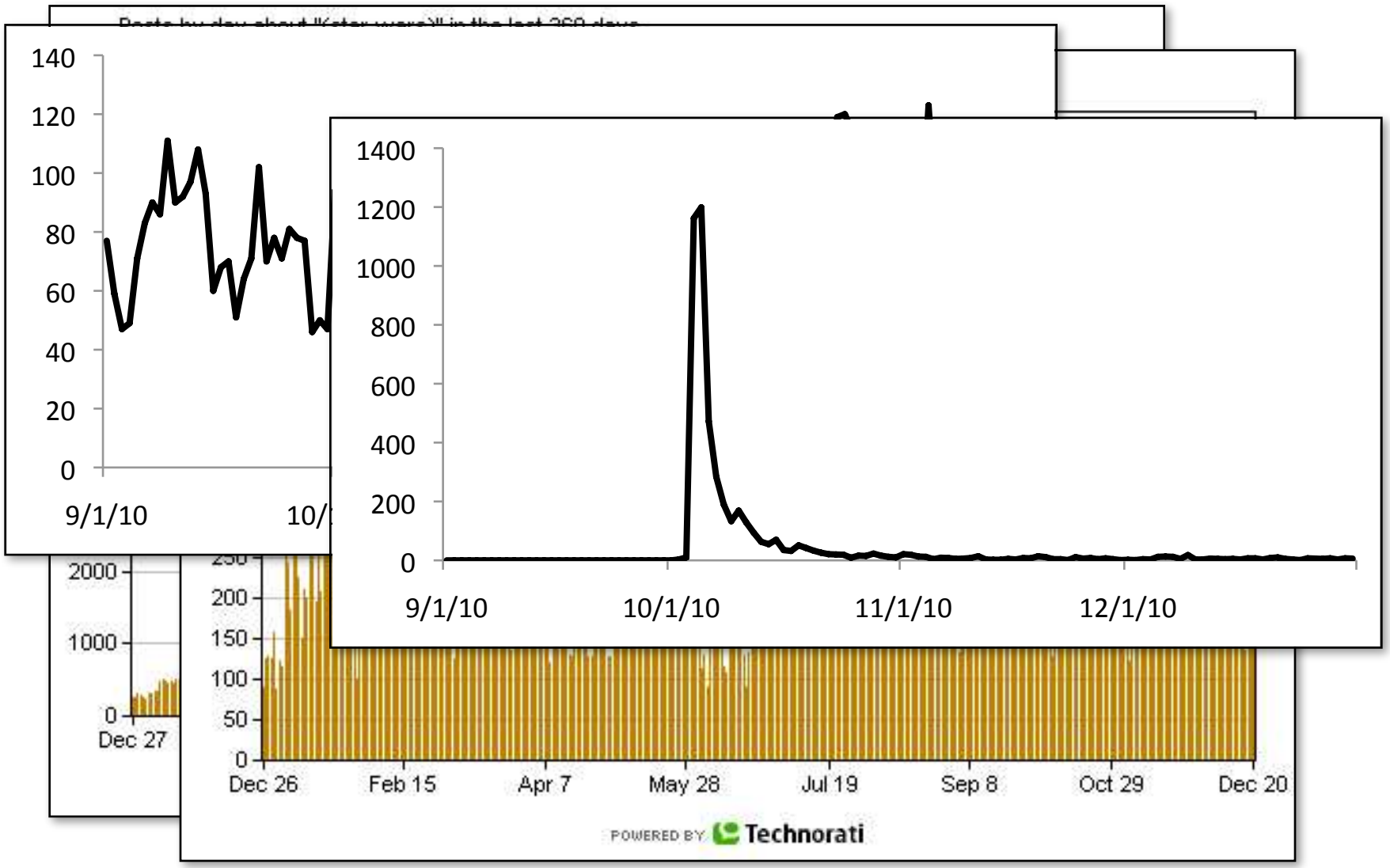
25

UNIVERSITEIT VAN AMSTERDAM



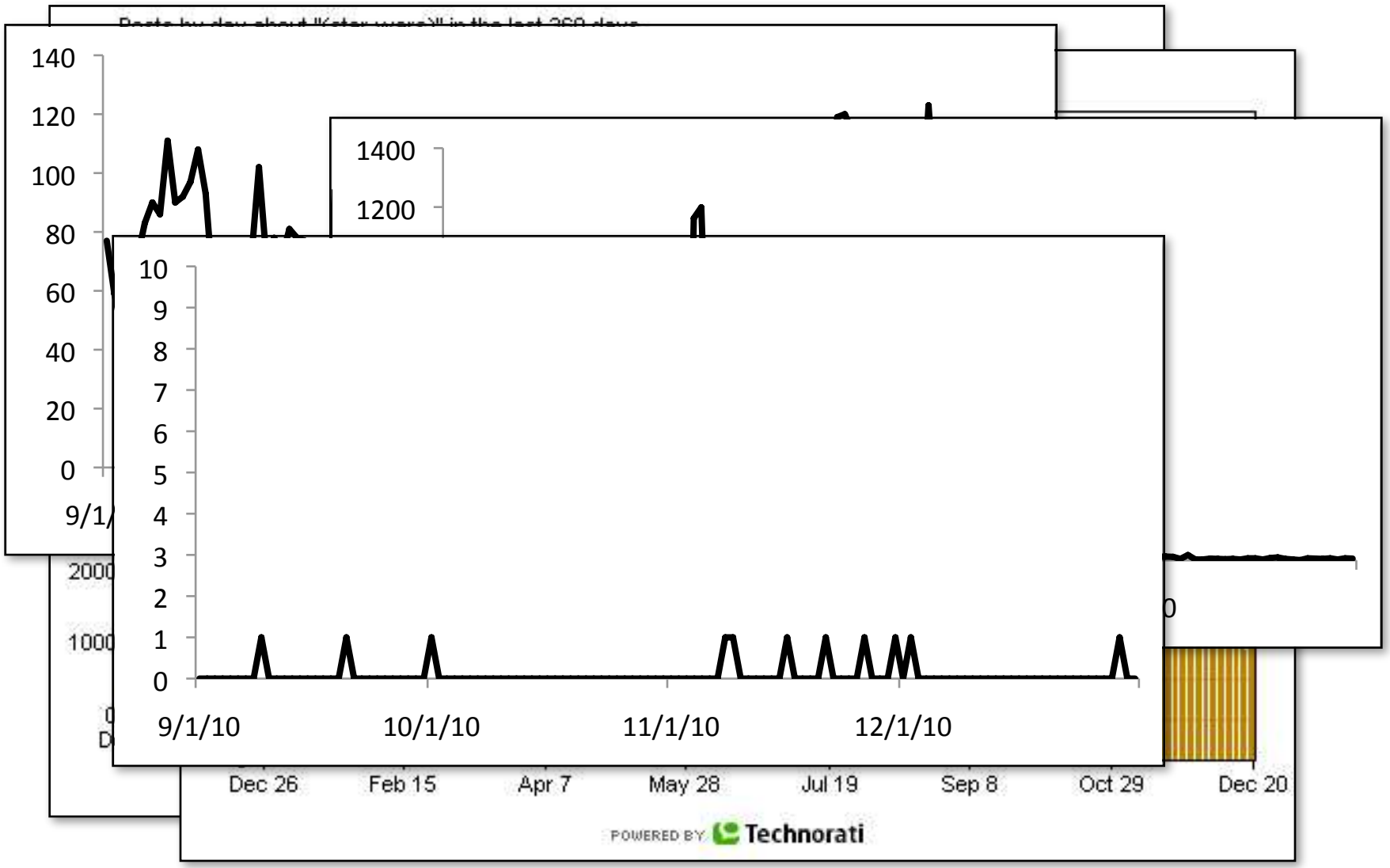
25

UNIVERSITEIT VAN AMSTERDAM



25

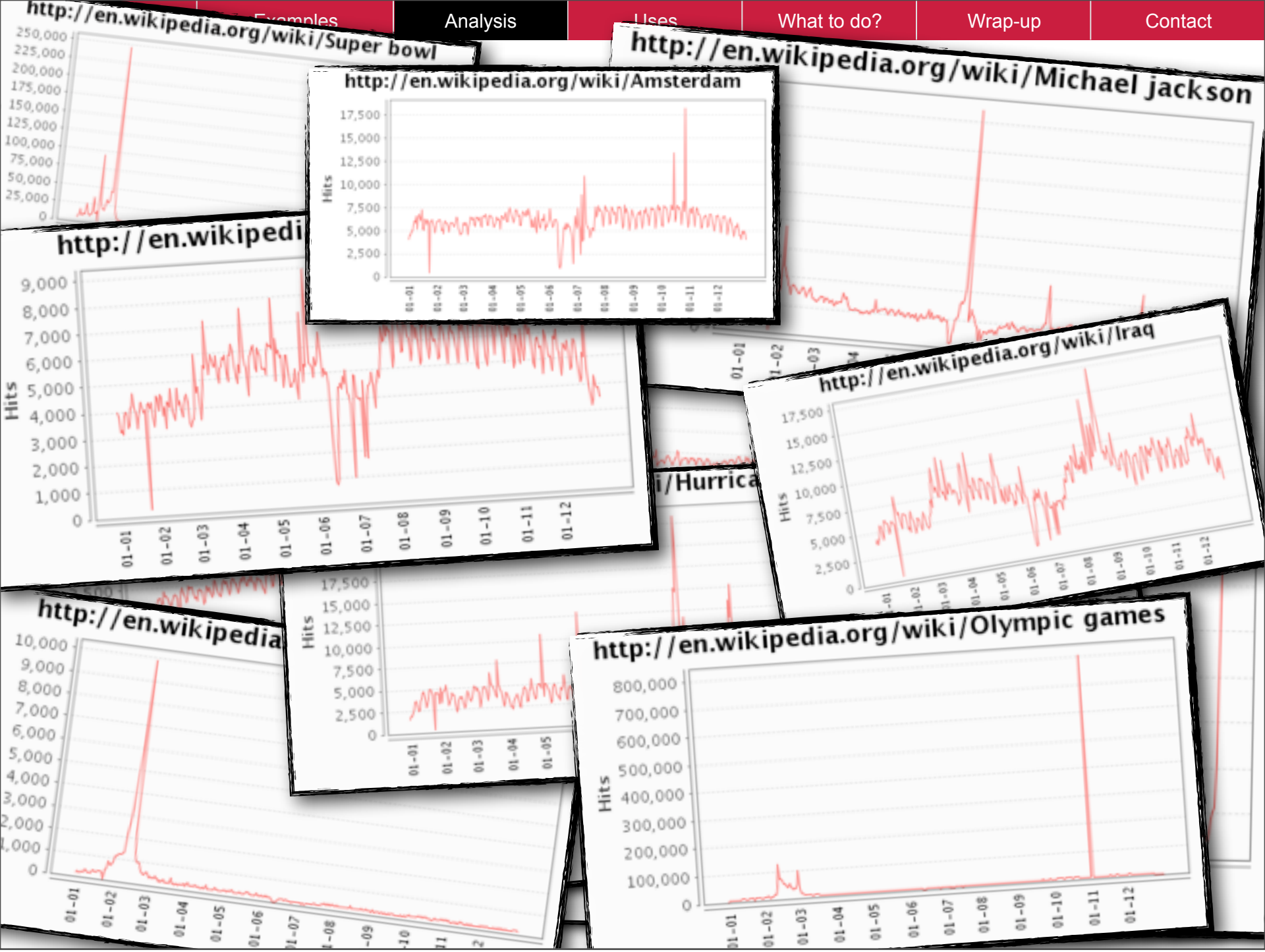
UNIVERSITEIT VAN AMSTERDAM



25

UNIVERSITEIT VAN AMSTERDAM

26


http://en.wikipedia.org/wiki/Super bowl
http://en.wikipedia.org/wiki/Amsterdam
http://en.wikipedia.org/wiki/Michael jackson
http://en.wikipedi
http://en.wikipedia.org/wiki/Iraq
http://en.wikipedia.org/wiki/Olympic games
Hits

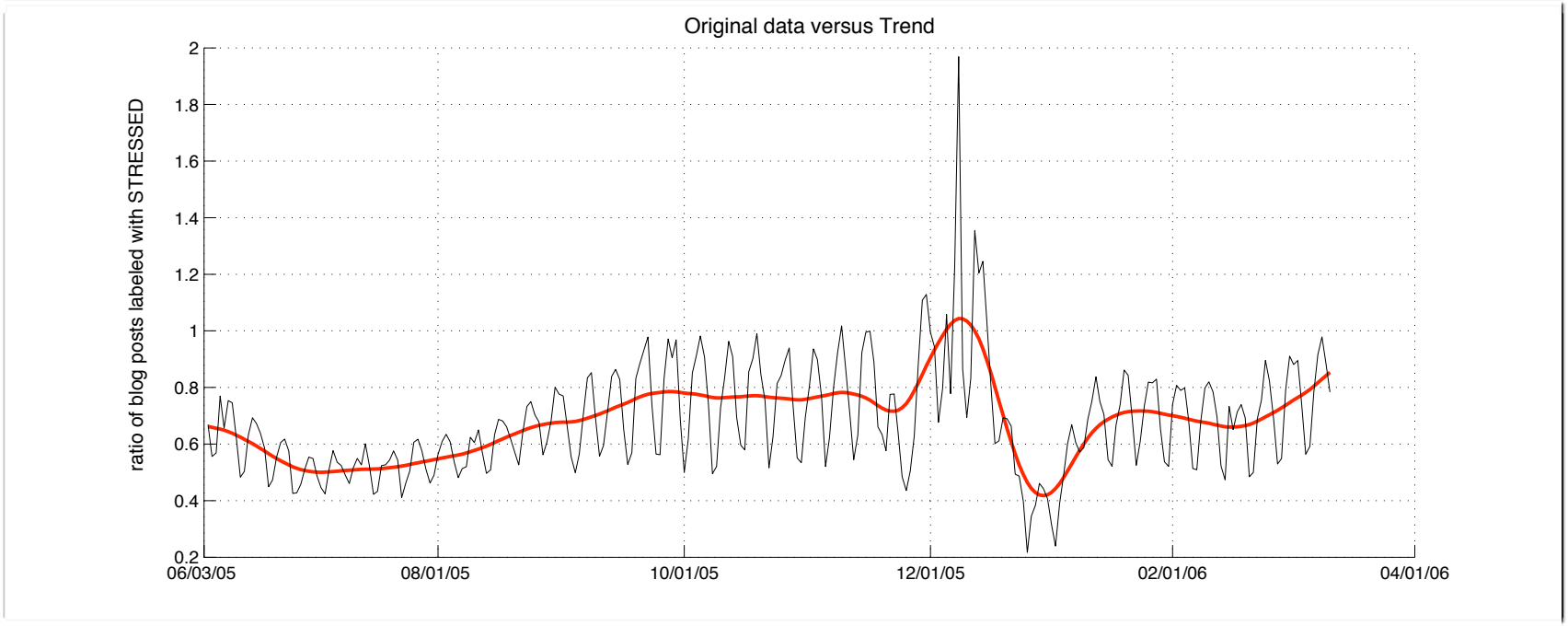
An aside …

27

UNIVERSITEIT VAN AMSTERDAM

An aside …

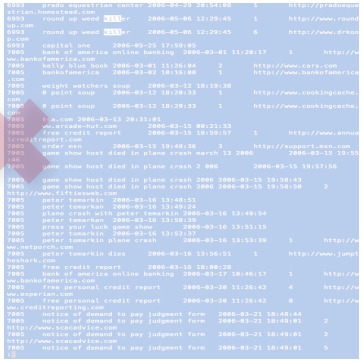
Tracking millions of bloggers and their moods



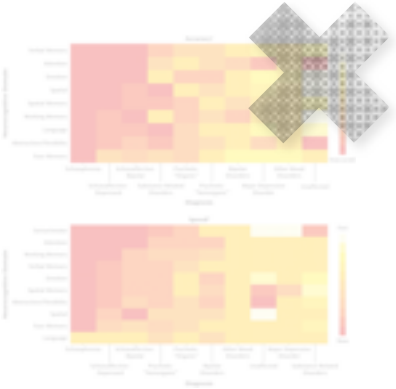
27

UNIVERSITEIT VAN AMSTERDAM

What

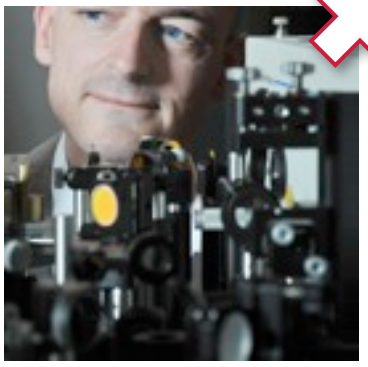


Examples



Analysis

Uses



What to do?

Wrap-up

Uses of log files

- Evaluation
- Query expansion
- Query suggestion
- Simulation and building models of user behavior
- Learning to rank
- Interleaved comparison

29

UNIVERSITEIT VAN AMSTERDAM

Evaluation

Evaluation

- Inform evaluation activities about generating test queries
  - TREC blog track
  - TREC session track
  - Invent information need underlying the query
- Is a click a judgment?
  - Traditionally, IR experiments use explicit relevance judgements.
  - Annotators examine queries and candidate documents, explicitly judging documents for relevance.
  - Use of explicit judgments is problematic
    - judging process takes a lot of time
    - there can be wide interannotator variation (Harter, 1996)
    - explicit judging may not result in the same assessments that a user would generate (Ruthven, 2005)
  - What about using click data from transaction logs to infer judgments?
    - (Joachims et al., 2005; Radlinski et al., 2008)

UNIVERSITEIT VAN AMSTERDAM

Evaluation (2)

Evaluation (2)

- Kamps et al. (2009) found large differences between system rankings based on explicit relevance assessments and those based on click data
- In a commercial setting, clicks and relevance may correlate very strongly (Hofmann et al, 2010)

Evaluation (2)

- Kamps et al. (2009) found large differences between system rankings based on explicit relevance assessments and those based on click data
- In a commercial setting, clicks and relevance may correlate very strongly (Hofmann et al, 2010)

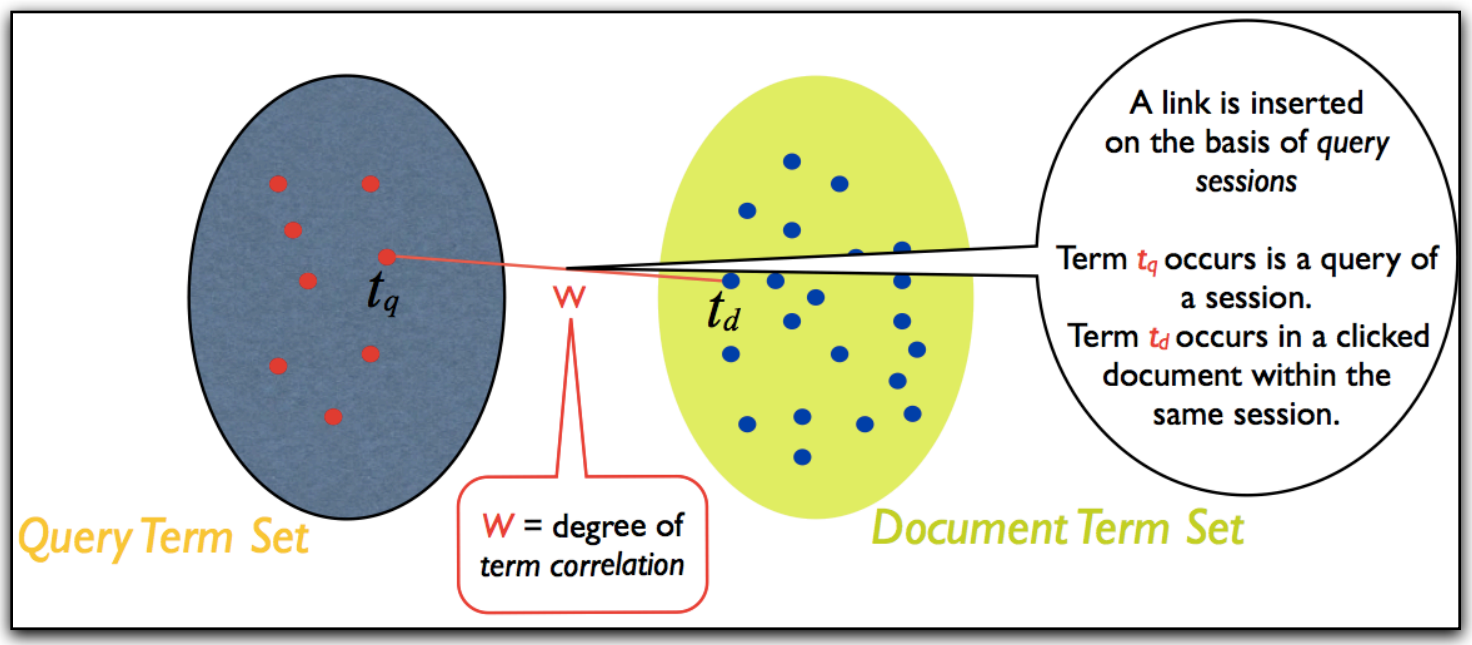
Table 1: Agreement between system rankings generated by *click* vs. *purchase* data according to standard evaluation measures. Agreement is calculated using Kendall's τ, and the number of pair-wise switches between ranked systems. All correlations are statistically significant with $p \ll 0.001$.

	purchases vs clicks from purchase queries		purchases vs clicks from non-purchase queries	
measure	τ	switches	τ	switches
MAP	0.974	6	0.766	54
MRR	0.948	12	0.766	54
P@10	0.991	2	0.775	52

31

Query expansion

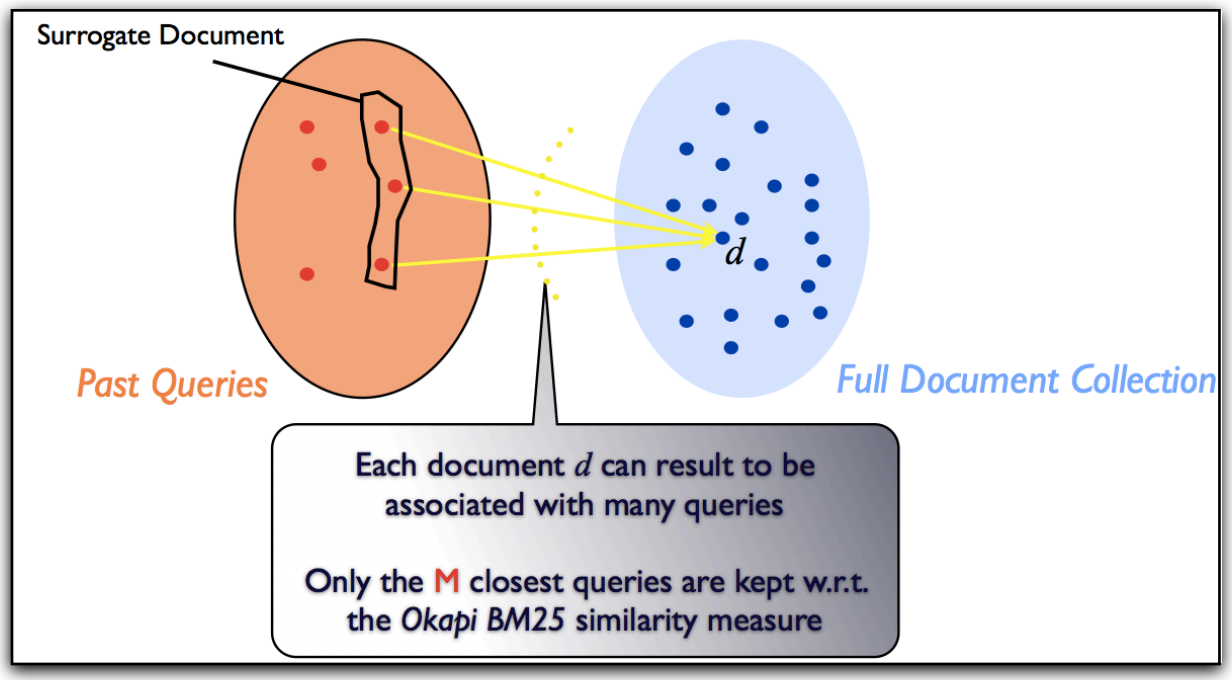
- Use correlations between terms in clicked documents and web search engine queries for query expansion
  - Cui et al, 2002
  - Can be used as a source of expansion terms



32

Query expansion (2)

- Use query-clicked document information to generate alternative representations of the documents
  - Billerbeck et al, 2003
  - Can be used as a source for expansion terms



33

Query suggestion

- Exploit information on past users' queries
  - Propose to a list of queries related to the one (or the ones, considering past queries in the same session) submitted
  - Query suggestion vs. expansion
    - users can select best similar query to refine search instead of having query uncontrollably stuffed with many terms
- Proposals in the literature
  - Queries suggested from those appearing frequently in query sessions
  - Use clustering to devise similar queries on the basis of cluster membership
  - Use click-through data information to devise query similarity

Query suggestion (2)

35

Query suggestion (2)

- Proposal 1 (Fonseca et al, 2003)
  - If a lot of users, after submitting query also submit query , then is a good suggestion for .
  - Can be solved as an association rule mining problem
- Proposal 2 (Baeza-Yates et al, 2004)
  - Two stages:
    - Offline: Create clusters of past queries based on query text along with the text of clicked URLs.
    - Online: Recommends queries on the basis of the input query
      - Find best matching cluster, and within cluster find “best” query
        - Attractiveness in terms of number of clicks generated, similarity, popularity, …
- Proposal 3 (Craswell et al 2006)
  - Random walks on the query-click bipartite graph (queries and clicked URLs), where the edges are symmetric

35

Simulation and model building

36

Simulation and model building

- Models for query generation, models for query generation, models for simulating clicks, …

36

Simulation and model building

- Models for query generation, models for query generation, models for simulating clicks, …
- Click model simulates user interactions
  - E.g., users traverse result lists from top to bottom
  - For each document they encounter, they decide whether the document representation is promising enough to warrant a click
    - E.g., based on URL, title, document snippets, preview
  - If, after clicking on a different, the user's information need is satisfied (likely of the document was relevant), they stop browsing the result list. Otherwise, they continue examining the result list (likely if the document was not relevant)

Simulation and model building

- Models for query generation, models for query generation, models for simulating clicks, …
- Click model simulates user interactions
  - E.g., users traverse result lists from top to bottom
  - For each document they encounter, they decide whether the document representation is promising enough to warrant a click
    - E.g., based on URL, title, document snippets, preview
  - If, after clicking on a different, the user's information need is satisfied (likely of the document was relevant), they stop browsing the result list. Otherwise, they continue examining the result list (likely if the document was not relevant)
- Why is this interesting? Or useful?

Simulation and model building

- Models for query generation, models for query generation, models for simulating clicks, …
- Click model simulates user interactions
  - E.g., users traverse result lists from top to bottom
  - For each document they encounter, they decide whether the document representation is promising enough to warrant a click
    - E.g., based on URL, title, document snippets, preview
  - If, after clicking on a different, the user's information need is satisfied (likely of the document was relevant), they stop browsing the result list. Otherwise, they continue examining the result list (likely if the document was not relevant)
- Why is this interesting? Or useful?
- What parameters do you need to set?

Simulation and model building

- Models for query generation, models for query generation, models for simulating clicks, …
- Click model simulates user interactions
  - E.g., users traverse result lists from top to bottom
  - For each document they encounter, they decide whether the document representation is promising enough to warrant a click
    - E.g., based on URL, title, document snippets, preview
  - If, after clicking on a different, the user's information need is satisfied (likely of the document was relevant), they stop browsing the result list. Otherwise, they continue examining the result list (likely if the document was not relevant)
- Why is this interesting? Or useful?
- What parameters do you need to set?
- How do log files help you?

Learning to rank

- Compute a global model to assign relevance score to each result page
  - First select best features to be used to identify importance of a page
  - Then train a machine learning algorithm (classifier/predictor) using these features on a ranked subset (training set) to learn a model
  - **If a document receives a click it is considered relevant for the query it has answered**
  - If *f* is a ranking function, we can define its performance as the average rank of the click results (lower is better)

$$\text{Perf}(f) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{|D_i|}\sum_{j=1}^{|D_i|}\text{rank}(f, Q_i, D_{ij})$$

D_{ij} : j-th document clicked in answer to query Q_i

Queries: $Q_1,\ldots,Q_{|Q|}$

Learning to rank (2)

Learning to rank (2)

- Click on a result is not an unbiased estimator for its importance
  - There is a natural bias towards clicking on top ranked docs
  - What query log features could give an **unbiased estimate** of relevance?
- Implicit relevance feedback from click-through data
  - If a user clicks on the i-th result item for a query, she considers it to be more important than previous ones
  - Induce relative preferences
- How accurate is this **implicit** feedback compared to **explicit** feedback
  - Joachims et al. (2007) compared pairwise preferences generated from clicks to explicit relevance judgments (through a user study)
  - Up to 89.7% (±9.5) preferences from clicks agree with the direction of a strict judgment of a human assessor
  - “Last Click > Skip Above”

UNIVERSITEIT VAN AMSTERDAM

Interleaved comparison

Interleaved comparison

- Compare rankers using **life** user interactions (e.g., clicks) that naturally occur in retrieval systems

Interleaved comparison

- Compare rankers using **life** user interactions (e.g., clicks) that naturally occur in retrieval systems
- To compare two systems (repeat over many queries)
  - Generate an interleaved list for each query based on the two rankers
  - User's clicks on the interleaved list are attributed to each ranker based on how they ranked the clicked document
  - The ranker that obtains more clicks deemed superior
  - This simple model has bias and sensitivity issues, but effective refinements are known
    (Hofmann et al., 2011)

Interleaved comparison

- Compare rankers using **life** user interactions (e.g., clicks) that naturally occur in retrieval systems
- To compare two systems (repeat over many queries)
  - Generate an interleaved list for each query based on the two rankers
  - User's clicks on the interleaved list are attributed to each ranker based on how they ranked the clicked document
  - The ranker that obtains more clicks deemed superior
  - This simple model has bias and sensitivity issues, but effective refinements are known (Hofmann et al., 2011)

Table 3: Accuracy for the comparison methods on ranker pairs constructed from individual features (after 1,000 queries).

run	*accuracy*
balanced interleave	0.881
team draft	0.898
document constraint	0.857
marginalized probabilities	**0.914**

Interleaved comparison

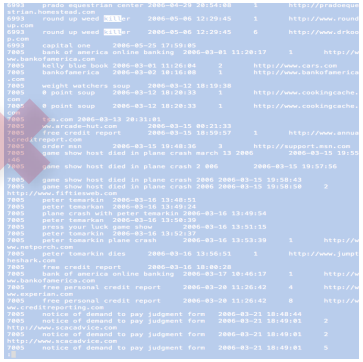
- Compare rankers using **life** user interactions (e.g., clicks) that naturally occur in retrieval systems
- To compare two systems (repeat over many queries)
  - Generate an interleaved list for each query based on the two rankers
  - User's clicks on the interleaved list are attributed to each ranker based on how they ranked the clicked document
  - The ranker that obtains more clicks deemed superior
  - This simple model has bias and sensitivity issues, but effective refinements are known (Hofmann et al., 2011)
- Transfer? Re-use?
  - "Logs again"

Table 3: Accuracy for the comparison methods on ranker pairs constructed from individual features (after 1,000 queries).

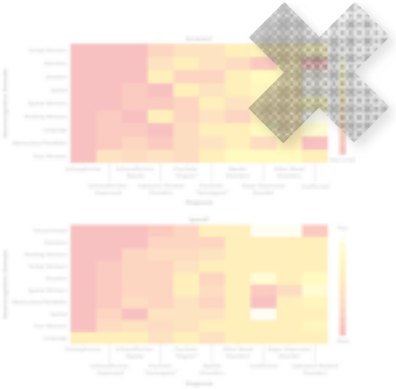
run	*accuracy*
balanced interleave	0.881
team draft	0.898
document constraint	0.857
marginalized probabilities	**0.914**

UNIVERSITEIT VAN AMSTERDAM

What



Examples



Analysis

Uses

What to do?

Wrap-up

Limitations of transaction log analysis

- Not all aspects of the search can be monitored
- Underlying information need (Rice and Borgman, 1983)
- Difficult to compare across transaction log studies of different systems due to system dependencies and varying implementations of analytical methods
- Lack of publicly available logs
  - Experiments not reproducible?
- Privacy
  - Queries, clicks, return visits, …, purchases, bookmarking, recommendations to friends, your whole life

41

UNIVERSITEIT VAN AMSTERDAM

How to get them

42

How to get them

- Generate them yourself
  - Build your own search engine for a niche
  - **Check with your university's ethics panel**

42

How to get them

- Generate them yourself
  - Build your own search engine for a niche
  - **Check with your university's ethics panel**
- Use open sources
  - RSS feeds, Wikipedia stats, …

42

How to get them

- Generate them yourself
  - Build your own search engine for a niche
  - **Check with your university's ethics panel**
- Use open sources
  - RSS feeds, Wikipedia stats, …
- Use your network
  - Convince organizations that do not have them yet, that they should get them and then help make sense of them
  - If the data cannot come to you, you should go to the data
    - Visit someone who has them
    - Do an internship with an organization that generates them
      - You know how they are

42

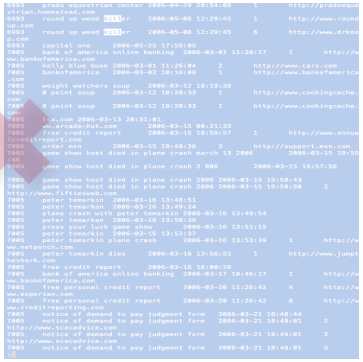
How to get them

- Generate them yourself
  - Build your own search engine for a niche
  - **Check with your university's ethics panel**
- Use open sources
  - RSS feeds, Wikipedia stats, …
- Use your network
  - Convince organizations that do not have them yet, that they should get them and then help make sense of them
  - If the data cannot come to you, you should go to the data
    - Visit someone who has them
    - Do an internship with an organization that generates them
      - You know how they are
- Build simulators — a perfectly valid and respectable way of testing your theories
  - But maybe not systems (?)

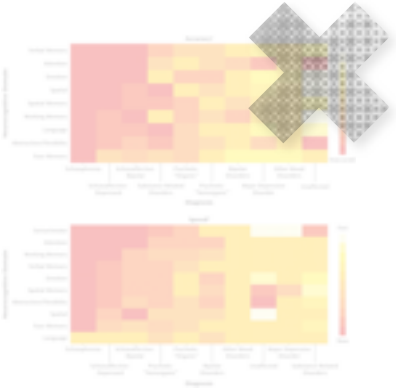
42

UNIVERSITEIT VAN AMSTERDAM

What



Examples



Analysis

Uses

What to do?

Wrap-up

UNIVERSITEIT VAN AMSTERDAM

Log file analysis

44

Log file analysis

- Descriptive aspects, different types of analysis, different types of uses
  - Very experimental, observational in nature
  - Results and techniques are diverse and fragmented

44

Log file analysis

- Descriptive aspects, different types of analysis, different types of uses
  - Very experimental, observational in nature
  - Results and techniques are diverse and fragmented

- Why bother with log file analyse?
  - **It is all about understanding user behavior at scale and without intruding**

44

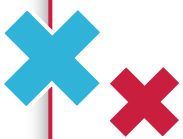
Log file analysis

- Descriptive aspects, different types of analysis, different types of uses
  - Very experimental, observational in nature
  - Results and techniques are diverse and fragmented
- Why bother with log file analyse?
  - **It is all about understanding user behavior at scale and without intruding**
- Why bother with user behavior?
  - Search is about user behavior
  - Search is getting ever more complex
    - Traditional IR
    - Structure (link, document, data)
    - User behavior

44

UNIVERSITEIT VAN AMSTERDAM

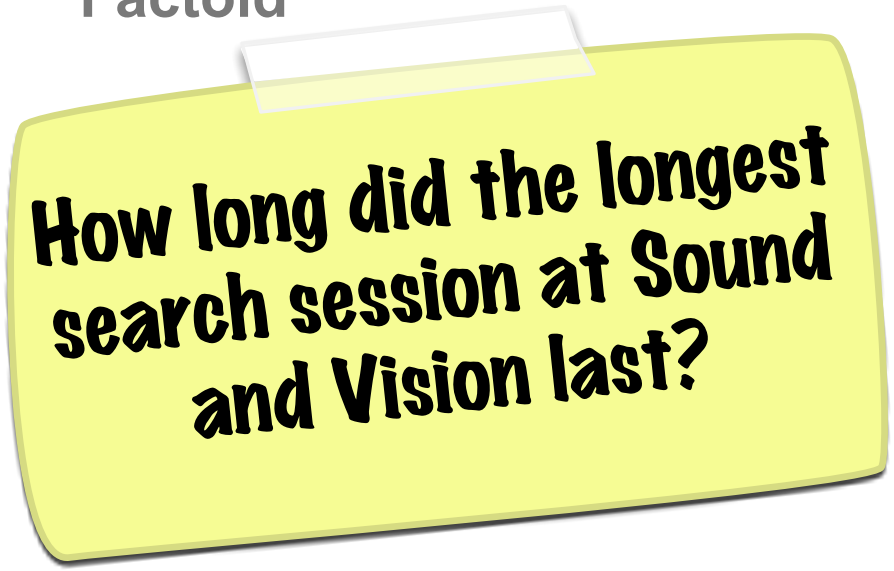
- Log File Analysis
- Maarten de Rijke
- derijke@uva.nl



UNIVERSITEIT VAN AMSTERDAM

Factoid

Factoid


How long did the longest search session at Sound and Vision last?

UNIVERSITEIT VAN AMSTERDAM

Factoid

How long did the longest search session at Sound and Vision last?

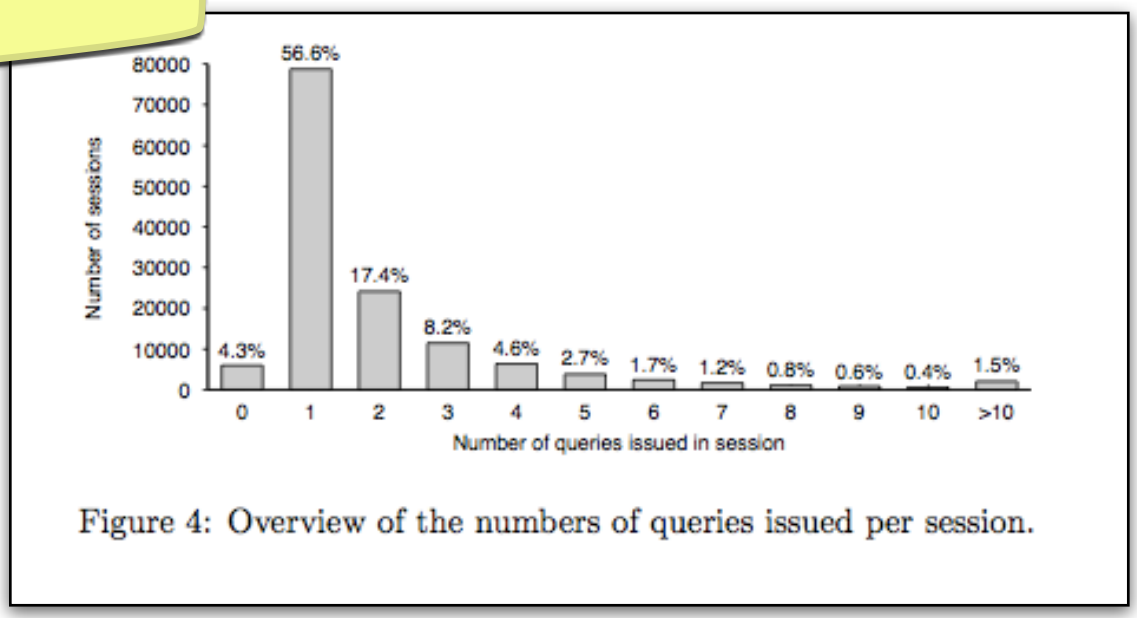


Figure 4: Overview of the numbers of queries issued per session.

46

UNIVERSITEIT VAN AMSTERDAM

Factoid

How long did the longest search session at Sound and Vision la

816:24:10
(hh:mm:ss)

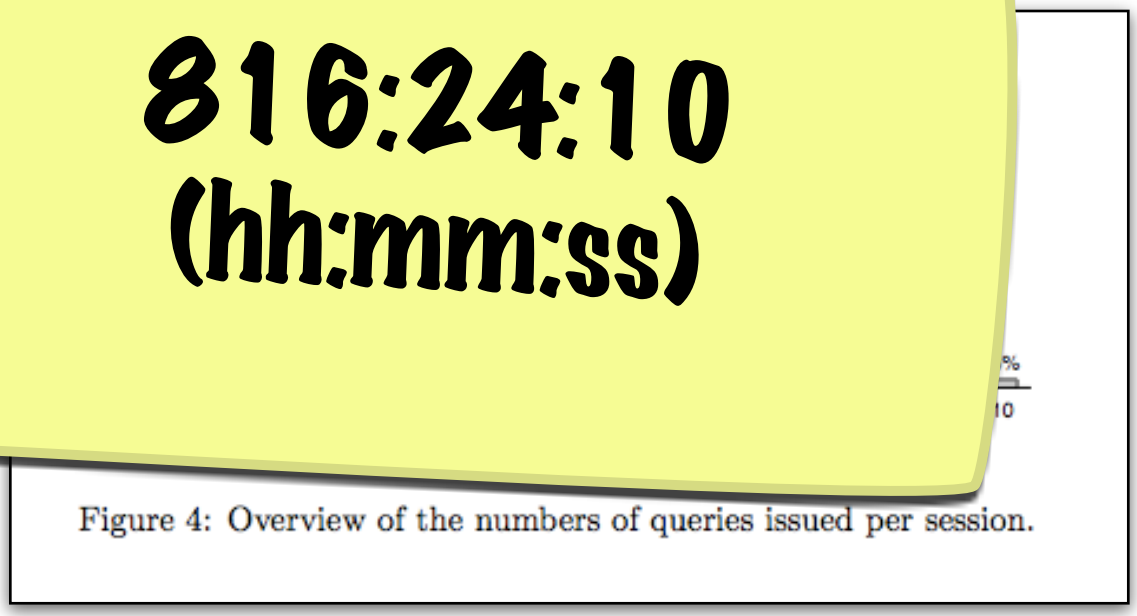


Figure 4: Overview of the numbers of queries issued per session.

46

UNIVERSITEIT VAN AMSTERDAM

Factoid

How

sear

Was it a successful session?

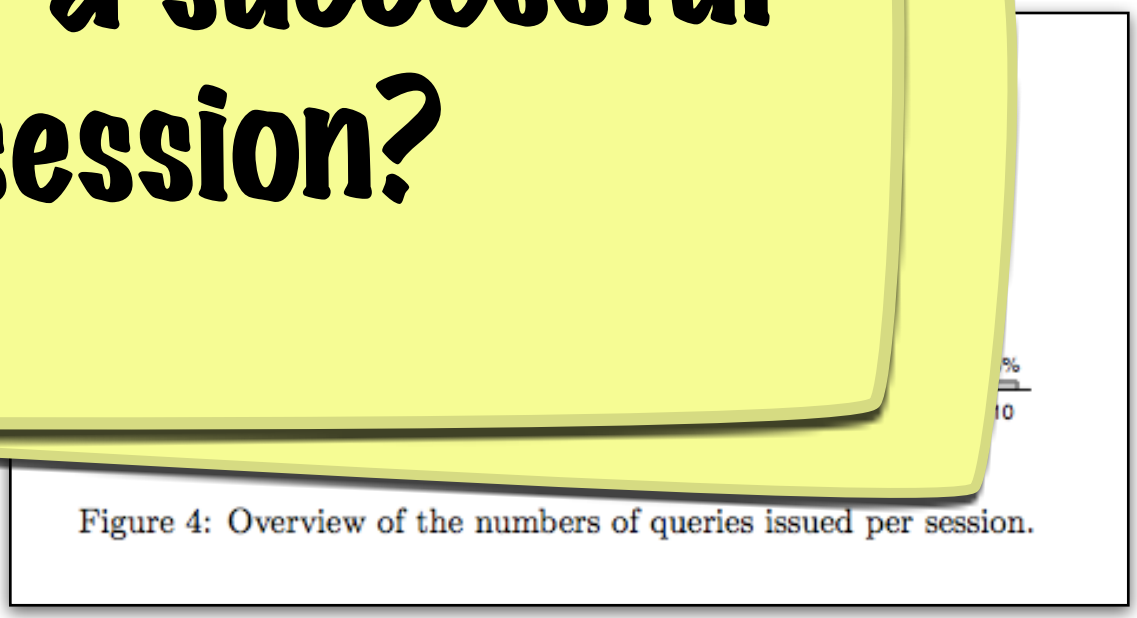


Figure 4: Overview of the numbers of queries issued per session.

46

Commercials

- http://elias-network.eu

- http://sigir.org/sigir2012/callfordoctoralconsortium.php

47