CE306/CE706
Information Retrieval

Search Log Analysis

Spring 2018

Brief Module Outline (Reminder)

- Motivation + introduction
- Processing pipeline / indexing + query processing
- Large-scale open-source search tools
- Information Retrieval models
- Evaluation
- Log analysis
- User profiles / personalisation / contextualisation
- IR applications, e.g. enterprise search

Log Analysis

- There are different types of log files, e.g.
  - Search logs (Web search, site search, …)
  - Server logs
  - Error logs
  - …
- We will be looking at the log files of a search engine
- This week we will focus on *Web search logs*
- Note: there is more than just *queries* in these logs …

Search Log Analysis

- Why is search log analysis important?
- What does a search log look like?
- Using search logs to better understand short- and long-term search tasks
- Using search logs to infer document relevance and ranking mistakes

Types of Evaluation (Reminder)

- Batch evaluation (using test collections)
- User studies
- Online evaluation

Test Collection-based Evaluation
advantages

- The experimental set-up is fixed: same queries, same corpus, same judgements
- Evaluations are reproducible: keeps us honest and allows us to easily measure improvement
- Modifying the system and re-evaluating is easy and free!
- A good way to tune parameters
- Makes error-analysis possible

Test Collection-based Evaluation
disadvantages

- Test-collection-building is time and resource intensive
- Human assessors are not users
- Makes assumptions that do not hold true in “real” life:
  - relevance is topical
  - context-independent
  - user-independent
  - stable over time

User Study Evaluation
advantages

- Can collect lots of data about users' reactions to a system
- The experimenter can control or manipulate the search task and the searcher's internal/external context
- Can collect lots of information about search outcomes
- Can be used to study unique populations of users

User-Study Evaluation
disadvantages

- Time and resource intensive, not a particularly good way to tune parameters
- The laboratory setting is not the user's normal environment
- Study participants know they are being 'observed'
- Not a good way to determine the frequency of natural events (especially rare ones)

Search-Log Analysis
general idea

- Can we reason about how well the system is performing by analyzing the search log?
- Can we use search-log information to improve its performance?
- Can we use search-log information to provide new services that enhance the user experience?
- All this goes beyond simply recording what a user is searching for, how long an average query is, how much time users spend on a search engine etc.

What is a Search-Log?

- Most search engines save information about every search
  - the query
  - a time-stamp
  - the IP address of the search client
  - the user id (stored in a cookie)
  - information about the search client (OS, browser, etc.)
  - the results that are presented
  - the results that are clicked
  - dwell time on a clicked result
  - ....

What is a Search-Log?

- This information is very sensitive and very valuable
- There are few publicly available Web search query-logs
  - the Excite Log (1997): ~18K users, ~50K queries
  - the AOL Log (2006): 650K users, ~20M queries
- Why aren't more search logs publicly available?
  - competitive reasons
  - privacy reasons
- In fact, the AOL Log was withdrawn and is no longer publicly available

What is a Search-Log?

SECTIONS HOME SEARCH The New York Times

TECHNOLOGY

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr. AUG. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Search-Logs and Privacy

- It is surprisingly easy to identify a person based on their queries
- Users prefer to remain anonymous
- We issue lots of “interesting” queries:
  - “how to tell a fake rolax”
  - “pictures of stars in the solar system”
  - “effective ways to fish a lizard”
  - “why does my iguana bob its head”

(AOL query-log)

What does a Search-Log Look Like?

```
...
903779;guest;83.33.xxx.xxx;83et8b7j010eh4vlht3ucj8dl1;en;
       ("pomegranate fertilization");search_sim;;0;-;;;2007-10-05 13:52:30
...
1889115;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;
        ("mozart");search_url;;0;-;;;2008-06-24 22:02:52
...
1889118;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;
        ("mozart");view_full;;1;;;;2008-06-24 22:03:03
...
1889120;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;
        Klavierkonzerte;search_res_rec_all;;0;-;;;2008-06-24 22:03:55
1889121;guest;71.249.xxx.xxx;8eb3bdv3odg9jncd71u0s2aff6;en;
        ("klavierkonzerte");view_full;;1;;;;2008-06-24 22:04:10
...
```

The European Library (TEL) logs

What does a Search-Log Look Like?

..
(1) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:17:41 +0200]
"GET /dbs.js HTTP/1.1" 200 236 http://www.bildungsserver.de/zeigen.html?seite=641 "Mozilla/5.0
(Windows; U; Windows NT 5.1; de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"
(2) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:17:41 +0200]
"GET /zeigen.html?seite=641 HTTP/1.1" 200 46606 http://www.google.de/search?client=firefox-
a&rls=org.mozilla%3Ade%3Aofficial&channel=s&hl=de&source=hp&q=schulamt&meta=&btnG=Google-
Suche "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"
(3) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:17:54 +0200]
"GET /metasuche/qsuche.html?feldinhalt1=**schulpflicht**&bool1=AND&finden=finden&searchall=ja&daten
banken%5B%5D=dbs_seiten&DBS=1&art=einfach HTTP/1.1" 200 139848
http://www.bildungsserver.de/zeigen.html?seite=641 "Mozilla/5.0 (Windows; U; Windows NT
5.1; de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"
(4) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:17:55 +0200]
"GET /metasuche/dbs.js HTTP/1.1" 200 236 http://www.bildungsserver.de/metasuche/qsuche.html?
feldinhalt1=schulpflicht&bool1=AND&finden=finden&searchall=ja&datenbanken%5B%5D=dbs_seiten
&DBS=1&art=einfach "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.14) Gecko/2009082707
Firefox/3.0.14"
(5) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:18:10 +0200]
"GET /metasuche/qsuche.html?feldinhalt1=**schulpflichtiger+Kinder** &bool1=AND&finden=finden
&searchall=ja&datenbanken%5B%5D=dbs_seiten&DBS=1&art=einfach HTTP/1.1" 200
109350 http://www.bildungsserver.de/metasuche/qsuche.html?feldinhalt1=schulpflicht
&bool1=AND&finden=finden&searchall=ja&datenbanken%5B%5D=dbs_seiten&DBS=1&art=einfach
"Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"
(6) bc76a65ad4be44e158b2d9ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:18:45 +0200]
"GET /zeigen.html?seite=21 HTTP/1.1" 200 25648 http://www.bildungsserver.de/metasuche/qsuche.html?
feldinhalt1=schulpflichtiger+Kinder&bool1=AND&finden=finden&searchall=ja &daten-
banken%5B%5D=dbs_seiten&DBS=1&art=einfach "Mozilla/5.0 (Windows; U; Windows NT 5.1;
de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"
(7) bc76a65ad4be44e158b2d9 ad17674f3dd4fa7296.rwth-aachen.de - - [23/Oct/2009:09:19:24 +0200]
"GET /zeigen.html?seite=136 HTTP/1.1" 200 45273 http://www.bildungsserver.de/zeigen.html?seite=21
"Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.14) Gecko/2009082707 Firefox/3.0.14"

..

Deutscher Bildungsserver (DBS) logs
(only entries (3) and (5) are actual search requests)

What does a Search-Log Look Like?

1024071	taraji henson	2006-03-02 00:28:45	4	http://www.tv.com
1024071	taraji henson	2006-03-02 00:28:45	1	http://www.imdb.com
1024071	the flavor of love vh1	2006-03-02 00:31:01	1	http://www.vh1.com
1024071	the flavor of love hoops	2006-03-02 00:38:32	1	http://www.vh1realityworld.com
1024071	beyonce	2006-03-02 22:42:05	1	http://www.beyonceonline.com
1024071	beyonce	2006-03-02 22:42:05	6	http://www.imdb.com
1024071	afc fighting	2006-03-04 22:35:33	2	http://sfuk.tripod.com
1024071	din thomas march 4th	2006-03-05 23:38:54	1	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	3	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	9	http://man-magazine.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	nit free picks	2006-03-15 14:02:21	1	http://www.docsports.com
1024071	1490 am radio	2006-03-15 14:48:01	8	http://www.1490wwpr.com
1024071	1490 am radio fl	2006-03-15 14:50:08	2	http://www.ontheradio.net
1024071	benihanas	2006-03-16 17:27:25	1	http://www.benihana.com
1024071	2006 winter music fest miami fl	2006-03-22 00:35:20	1	http://www.wintermusicconference.com
1024071	hotmail	2006-04-01 18:49:02	1	http://www.hotmail.com
1024071	my space	2006-04-02 01:21:41	1	http://www.myspace.com
1024071	my space	2006-04-02 15:59:20	1	http://www.myspace.com
1024071	my space	2006-04-02 22:03:10	1	http://www.myspace.com
1024071	nba jams super nintendo cheats	2006-04-03 21:06:11	2	http://www.elook.org
1024071	my space	2006-04-03 21:16:00	1	http://www.myspace.com
1024071	charlie's dodge fort pierce	2006-05-08 20:06:17	1	http://www.dealernet.com
1024071	charlie's dodge of fort pierce used cars	2006-05-08 20:09:27	2	http://www.automotive.com
1024071	justin timberlake new album	2006-05-12 16:21:36	4	http://www.mtv.com
1024071	mike epps	2006-05-13 19:45:56	6	http://www.hollywood.com
1024071	mike epps bio	2006-05-13 19:51:05	4	http://movies.aol.com
1024071	mike epps bio	2006-05-13 19:51:05	9	http://www.moono.com
1024071	mike epps bio	2006-05-13 19:55:56	14	http://video.barnesandnoble.com
1024071	mike epps bio	2006-05-13 19:55:56	21	http://www.hbo.com
1024071	mike epps bio	2006-05-13 20:01:06	24	http://www.vh1.com
1024071	mind freak	2006-05-14 00:46:18	10	http://video.google.com
1024071	criss angel mind freak	2006-05-14 12:53:35	1	http://www.crissangel.com
1024071	criss angel mind freak	2006-05-14 12:53:35	8	http://www.imdb.com
1024071	06-06-06	2006-05-14 22:29:11	1	http://www.timesonline.co.uk
1024071	show and sell auto fort pierce fl	2006-05-15 16:58:53	1	http://www.traderonline.com
1024071	barry bonds homerun ball 714 for sale	2006-05-25 16:25:41	5	http://www.sportsnet.ca
1024071	ufc 60 live results	2006-05-27 23:00:38	4	http://www.prowrestling.com
1024071	ufc 60 live play by play	2006-05-27 23:07:16	4	http://www.24wrestling.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	1	http://www.aplusmodel.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	8	http://www.inc.com

(AOL query-log)

What does a Search-Log Look Like?

1024071	taraji henson	2006-03-02 00:28:45	4 http://www.tv.com
1024071	taraji henson	2006-03-02 00:28:45	1 http://www.imdb.com
1024071	the flavor of love vh1	2006-03-02 00:31:01	1 http://www.vh1.com
1024071	the flavor of love hoops	2006-03-02 00:38:32	1 http://www.vh1realityworld.com
1024071	beyonce	2006-03-02 22:42:05	1 http://www.beyonceonline.com
1024071	beyonce	2006-03-02 22:42:05	6 http://www.imdb.com
1024071	afc fighting	2006-03-04 22:35:33	2 http://sfuk.tripod.com
1024071	din thomas march 4th	2006-03-05 23:38:54	1 http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	3 http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	9 http://man-magazine.com
1024071	unc basketball roster	2006-03-09 23:45:15	2 http://tarheelblue.collegesports.com
1024071	unc basketball roster	2006-03-09 23:45:15	2 http://tarheelblue.collegesports.com
1024071	nit free picks	2006-03-15 14:02:21	1 http://www.docsports.com
1024071	1490 am radio	2006-03-15 14:48:01	8 http://www.1490wwpr.com
1024071	1490 am radio fl	2006-03-15 14:50:08	2 http://www.ontheradio.net
1024071	benihanas	2006-03-16 17:27:25	1 http://www.benihana.com
1024071	2006 winter music fest miami fl	2006-03-22 00:35:20	1 http://www.wintermusicconference.com
1024071	hotmail	2006-04-01 18:49:02	1 http://www.hotmail.com
1024071	my space	2006-04-02 01:21:41	1 http://www.myspace.com
1024071	my space	2006-04-02 15:59:20	1 http://www.myspace.com
1024071	my space	2006-04-02 22:03:10	1 http://www.myspace.com
1024071	nba jams super nintendo cheats	2006-04-03 21:06:11	2 http://www.elook.org
1024071	my space	2006-04-03 21:16:00	1 http://www.myspace.com
1024071	charlie's dodge fort pierce	2006-05-08 20:06:17	1 http://www.dealernet.com
1024071	charlie's dodge of fort pierce used cars	2006-05-08 20:09:27	2 http://www.automotive.com
1024071	justin timberlake new album	2006-05-12 16:21:36	4 http://www.mtv.com
1024071	mike epps	2006-05-13 19:45:56	6 http://www.hollywood.com
1024071	mike epps bio	2006-05-13 19:51:05	4 http://movies.aol.com
1024071	mike epps bio	2006-05-13 19:51:05	9 http://www.moono.com
1024071	mike epps bio	2006-05-13 19:55:56	14 http://video.barnesandnoble.com
1024071	mike epps bio	2006-05-13 19:55:56	21 http://www.hbo.com
1024071	mike epps bio	2006-05-13 20:01:06	24 http://www.vh1.com
1024071	mind freak	2006-05-14 00:46:18	10 http://video.google.com
1024071	criss angel mind freak	2006-05-14 12:53:35	1 http://www.crissangel.com
1024071	criss angel mind freak	2006-05-14 12:53:35	8 http://www.imdb.com
1024071	06-06-06	2006-05-14 22:29:11	1 http://www.timesonline.co.uk
1024071	show and sell auto fort pierce fl	2006-05-15 16:58:53	1 http://www.traderonline.com
1024071	barry bonds homerun ball 714 for sale	2006-05-25 16:25:41	5 http://www.sportsnet.ca
1024071	ufc 60 live results	2006-05-27 23:00:38	4 http://www.prowrestling.com
1024071	ufc 60 live play by play	2006-05-27 23:07:16	4 http://www.24wrestling.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	1 http://www.aplusmodel.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	8 http://www.inc.com

what kinds of things could we do with this?

(AOL query-log)

Usefulness of Search-Logs

Rank	Query Phrase	Rank	Query Phrase
1	mozart	11	dante
2	harry potter	12	zagreb
3	meisje met de parel	13	bible
4	einstein	14	poland
5	shakespeare	15	history
6	bach	16	france
7	music	17	chopin
8	europe	18	paris
9	goethe	19	italy
10	london	20	cervantes

Frequent queries in TEL (for sample log)

Usefulness of Search-Logs

Count	Query
27133	moodle
16382	library
11624	timetable
10879	search
5510	cmr
4913	enrol
4543	*(empty query)*
3745	accommodation
3740	ocs
3711	accomodation
3565	graduation
3492	psychology
3381	timetables
2969	term dates
2769	courses
2704	student union
2310	fees
2241	law
2238	sports centre
2203	registry
2097	exam timetable
2058	mba

Frequent queries submitted to a university search engine (for sample log)

Usefulness of Search-Logs

- Spelling corrections
- Query suggestions
- Query expansion
- Query classification: informational, navigational, transactional
- Vertical selection and presentation
- Personalization
- Detecting commercial intent (ad placement)
- Predicting query ambiguity
- Evaluation
- Detecting ranking mistakes
- Inferring sub-tasks associated with query

What does a Search-Log Look Like?

1024071	taraji henson	2006-03-02 00:28:45	4	http://www.tv.com
1024071	taraji henson	2006-03-02 00:28:45	1	http://www.imdb.com
1024071	the flavor of love vh1	2006-03-02 00:31:01	1	http://www.vh1.com
1024071	the flavor of love hoops	2006-03-02 00:38:32	1	http://www.vh1realityworld.com
1024071	beyonce	2006-03-02 22:42:05	1	http://www.beyonceonline.com
1024071	beyonce	2006-03-02 22:42:05	6	http://www.imdb.com
1024071	afc fighting	2006-03-04 22:35:33	2	http://sfuk.tripod.com
1024071	din thomas march 4th	2006-03-05 23:38:54	1	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	3	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	9	http://man-magazine.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	nit free picks	2006-03-15 14:02:21	1	http://www.docsports.com
1024071	1490 am radio	2006-03-15 14:48:01	8	http://www.1490wwpr.com
1024071	1490 am radio fl	2006-03-15 14:50:08	2	http://www.ontheradio.net
1024071	benihanas	2006-03-16 17:27:25	1	http://www.benihana.com
1024071	2006 winter music fest miami fl	2006-03-22 00:35:20	1	http://www.wintermusicconference.com
1024071	hotmail	2006-04-01 18:49:02	1	http://www.hotmail.com
1024071	my space	2006-04-02 01:21:41	1	http://www.myspace.com
1024071	my space	2006-04-02 15:59:20	1	http://www.myspace.com
1024071	my space	2006-04-02 22:03:10	1	http://www.myspace.com
1024071	nba jams super nintendo cheats	2006-04-03 21:06:11	2	http://www.elook.org
1024071	my space	2006-04-03 21:16:00	1	http://www.myspace.com
1024071	charlie's dodge fort pierce	2006-05-08 20:06:17	1	http://www.dealernet.com
1024071	charlie's dodge of fort pierce used cars	2006-05-08 20:09:27	2	http://www.automotive.com
1024071	justin timberlake new album	2006-05-12 16:21:36	4	http://www.mtv.com
1024071	mike epps	2006-05-13 19:45:56	6	http://www.hollywood.com
1024071	mike epps bio	2006-05-13 19:51:05	4	http://movies.aol.com
1024071	mike epps bio	2006-05-13 19:51:05	9	http://www.moono.com
1024071	mike epps bio	2006-05-13 19:55:56	14	http://video.barnesandnoble.com
1024071	mike epps bio	2006-05-13 19:55:56	21	http://www.hbo.com
1024071	mike epps bio	2006-05-13 20:01:06	24	http://www.vh1.com
1024071	mind freak	2006-05-14 00:46:18	10	http://video.google.com
1024071	criss angel mind freak	2006-05-14 12:53:35	1	http://www.crissangel.com
1024071	criss angel mind freak	2006-05-14 12:53:35	8	http://www.imdb.com
1024071	06-06-06	2006-05-14 22:29:11	1	http://www.timesonline.co.uk
1024071	show and sell auto fort pierce fl	2006-05-15 16:58:53	1	http://www.traderonline.com
1024071	barry bonds homerun ball 714 for sale	2006-05-25 16:25:41	5	http://www.sportsnet.ca
1024071	ufc 60 live results	2006-05-27 23:00:38	4	http://www.prowrestling.com
1024071	ufc 60 live play by play	2006-05-27 23:07:16	4	http://www.24wrestling.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	1	http://www.aplusmodel.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	8	http://www.inc.com

are these queries independent?

(AOL query-log)

Search Sessions

- Search is a “dialogue” between a user and a search engine
  - user: query
  - search engine: search results
  - user: reformulated query
  - search engine: new search results
- Each “dialogue” is called a search session
- Each dialogue corresponds to an information need (at some level of granularity)
- A dialogue ends when the user is satisfied or gives up

Search Sessions

- Question: what proportion of search sessions result in user-satisfaction?
- The answer may be in the search log
- But, first, we have to recover each individual dialogue
- Requires some amount of "detective work" (easier in the TEL and DBS examples as they contain *session ids*)
- The simplest approaches assume that same-dialogue queries are sequential
- In other words, users engage in one dialogue at a time
- Are there environments where this is or is not a valid assumption?

Search Sessions

1024071	taraji henson	2006-03-02 00:28:45	4	http://www.tv.com
1024071	taraji henson	2006-03-02 00:28:45	1	http://www.imdb.com
1024071	the flavor of love vh1	2006-03-02 00:31:01	1	http://www.vh1.com
1024071	the flavor of love hoops	2006-03-02 00:38:32	1	http://www.vh1realityworld.com
1024071	beyonce	2006-03-02 22:42:05	1	http://www.beyonceonline.com
1024071	beyonce	2006-03-02 22:42:05	6	http://www.imdb.com
1024071	afc fighting	2006-03-04 22:35:33	2	http://sfuk.tripod.com
1024071	din thomas march 4th	2006-03-05 23:38:54	1	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	3	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	9	http://man-magazine.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	nit free picks	2006-03-15 14:02:21	1	http://www.docsports.com
1024071	1490 am radio	2006-03-15 14:48:01	8	http://www.1490wwpr.com
1024071	1490 am radio fl	2006-03-15 14:50:08	2	http://www.ontheradio.net
1024071	benihanas	2006-03-16 17:27:25	1	http://www.benihana.com
1024071	2006 winter music fest miami fl	2006-03-22 00:35:20	1	http://www.wintermusicconference.com
1024071	hotmail 2006-04-01 18:49:02	1	http://www.hotmail.com	
1024071	my space	2006-04-02 01:21:41	1	http://www.myspace.com
1024071	my space	2006-04-02 15:59:20	1	http://www.myspace.com
1024071	my space	2006-04-02 22:03:10	1	http://www.myspace.com
1024071	nba jams super nintendo cheats	2006-04-03 21:06:11	2	http://www.elook.org
1024071	my space	2006-04-03 21:16:00	1	http://www.myspace.com
1024071	charlie's dodge fort pierce	2006-05-08 20:06:17	1	http://www.dealernet.com
1024071	charlie's dodge of fort pierce used cars	2006-05-08 20:09:27	2	http://www.automotive.com
1024071	justin timberlake new album	2006-05-12 16:21:36	4	http://www.mtv.com
1024071	mike epps	2006-05-13 19:45:56	6	http://www.hollywood.com
1024071	mike epps bio	2006-05-13 19:51:05	4	http://movies.aol.com
1024071	mike epps bio	2006-05-13 19:51:05	9	http://www.moono.com
1024071	mike epps bio	2006-05-13 19:55:56	14	http://video.barnesandnoble.com
1024071	mike epps bio	2006-05-13 19:55:56	21	http://www.hbo.com
1024071	mike epps bio	2006-05-13 20:01:06	24	http://www.vh1.com
1024071	mind freak	2006-05-14 00:46:18	10	http://video.google.com
1024071	criss angel mind freak	2006-05-14 12:53:35	1	http://www.crissangel.com
1024071	criss angel mind freak	2006-05-14 12:53:35	8	http://www.imdb.com
1024071	06-06-06	2006-05-14 22:29:11	1	http://www.timesonline.co.uk
1024071	show and sell auto fort pierce fl	2006-05-15 16:58:53	1	http://www.traderonline.com
1024071	barry bonds homerun ball 714 for sale	2006-05-25 16:25:41	5	http://www.sportsnet.ca
1024071	ufc 60 live results	2006-05-27 23:00:38	4	http://www.prowrestling.com
1024071	ufc 60 live play by play	2006-05-27 23:07:16	4	http://www.24wrestling.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	1	http://www.aplusmodel.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	8	http://www.inc.com

(AOL query-log)

Search Sessions

1024071	taraji henson	2006-03-02 00:28:45	4	http://www.tv.com
1024071	taraji henson	2006-03-02 00:28:45	1	http://www.imdb.com
1024071	the flavor of love vh1	2006-03-02 00:31:01	1	http://www.vh1.com
1024071	the flavor of love hoops	2006-03-02 00:38:32	1	http://www.vh1realityworld.com
1024071	beyonce	2006-03-02 22:42:05	1	http://www.beyonceonline.com
1024071	beyonce	2006-03-02 22:42:05	6	http://www.imdb.com
1024071	afc fighting	2006-03-04 22:35:33	2	http://sfuk.tripod.com
1024071	din thomas march 4th	2006-03-05 23:38:54	1	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	3	http://www.mmaringreport.com
1024071	mfc march 4th results	2006-03-05 23:45:49	9	http://man-magazine.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	unc basketball roster	2006-03-09 23:45:15	2	http://tarheelblue.collegesports.com
1024071	nit free picks	2006-03-15 14:02:21	1	http://www.docsports.com
1024071	1490 am radio	2006-03-15 14:48:01	8	http://www.1490wwpr.com
1024071	1490 am radio fl	2006-03-15 14:50:08	2	http://www.ontheradio.net
1024071	benihanas	2006-03-16 17:27:25	1	http://www.benihana.com
1024071	2006 winter music fest miami fl	2006-03-22 00:35:20	1	http://www.wintermusicconference.com
1024071	hotmail	2006-04-01 18:49:02	1	http://www.hotmail.com
1024071	my space	2006-04-02 01:21:41	1	http://www.myspace.com
1024071	my space	2006-04-02 15:59:20	1	http://www.myspace.com
1024071	my space	2006-04-02 22:03:10	1	http://www.myspace.com
1024071	nba jams super nintendo cheats	2006-04-03 21:06:11	2	http://www.elook.org
1024071	my space	2006-04-03 21:16:00	1	http://www.myspace.com
1024071	charlie's dodge fort pierce	2006-05-08 20:06:17	1	http://www.dealernet.com
1024071	charlie's dodge of fort pierce used cars	2006-05-08 20:09:27	2	http://www.automotive.com
1024071	justin timberlake new album	2006-05-12 16:21:36	4	http://www.mtv.com

(AOL query-log)

Search Sessions

1024071	mike epps	2006-05-13 19:45:56	6	http://www.hollywood.com
1024071	mike epps bio	2006-05-13 19:51:05	4	http://movies.aol.com
1024071	mike epps bio	2006-05-13 19:51:05	9	http://www.moono.com
1024071	mike epps bio	2006-05-13 19:55:56	14	http://video.barnesandnoble.com
1024071	mike epps bio	2006-05-13 19:55:56	21	http://www.hbo.com
1024071	mike epps bio	2006-05-13 20:01:06	24	http://www.vh1.com
1024071	mind freak	2006-05-14 00:46:18	10	http://video.google.com
1024071	criss angel mind freak	2006-05-14 12:53:35	1	http://www.crissangel.com
1024071	criss angel mind freak	2006-05-14 12:53:35	8	http://www.imdb.com
1024071	06-06-06	2006-05-14 22:29:11	1	http://www.timesonline.co.uk
1024071	show and sell auto fort pierce fl	2006-05-15 16:58:53	1	http://www.traderonline.com
1024071	barry bonds homerun ball 714 for sale	2006-05-25 16:25:41	5	http://www.sportsnet.ca
1024071	ufc 60 live results	2006-05-27 23:00:38	4	http://www.prowrestling.com
1024071	ufc 60 live play by play	2006-05-27 23:07:16	4	http://www.24wrestling.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	1	http://www.aplusmodel.com
1024071	how to tell a fake rolax	2006-05-29 14:53:53	8	http://www.inc.com
1024071	locating serial number on rolex	2006-05-30 21:51:34	1	http://www.qualitytyme.net

(AOL query-log)

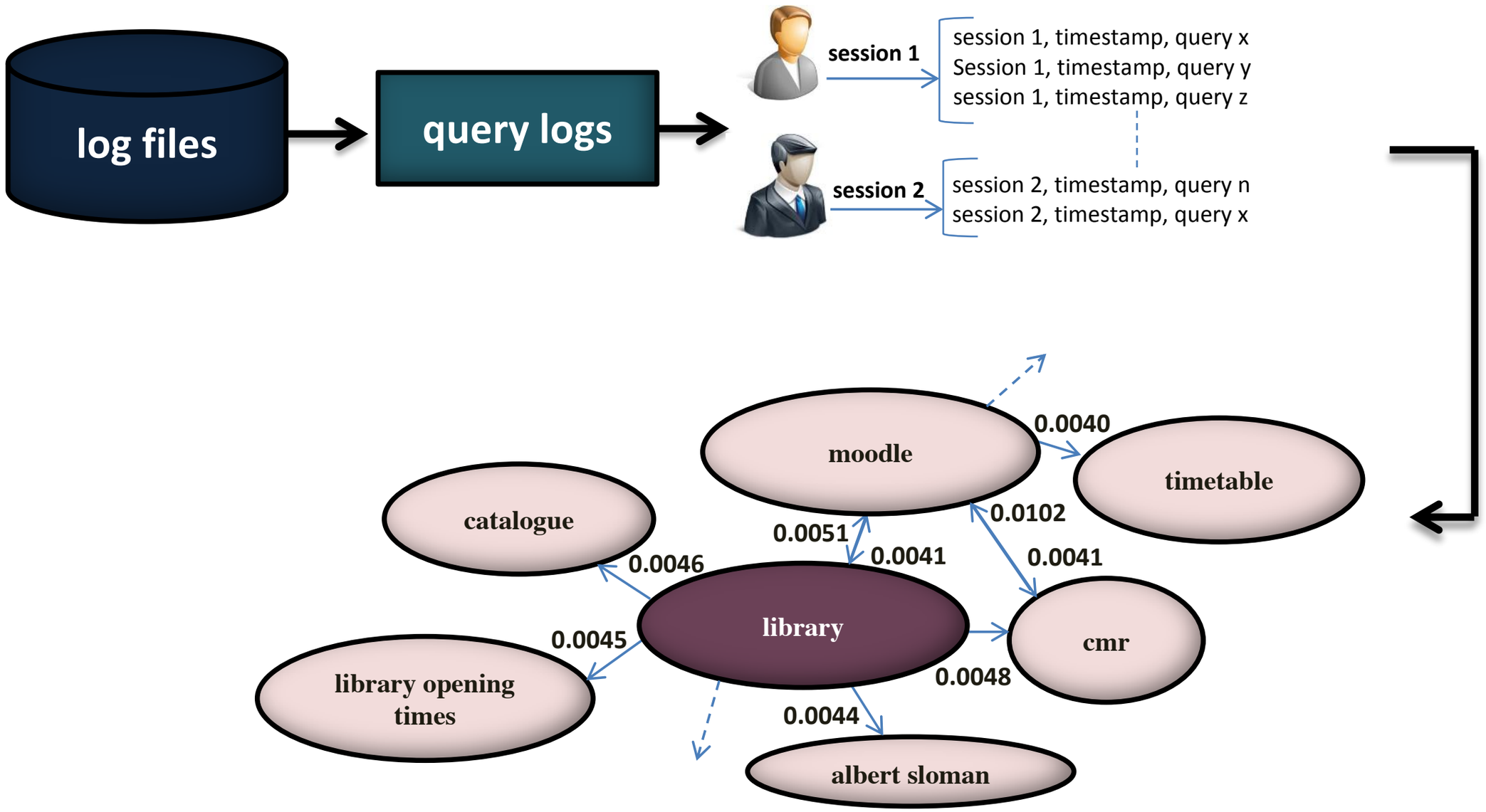
Heuristics for Recovering Search Sessions

- Time difference: subsequent queries are part of the same session if the difference between time-stamps is < t
  - 30 minutes is commonly used
  - Web servers tend to use 30 minutes of inactivity to assign a new session id (see TEL and DBS logs)
  - These are just approximations!
- Common term: subsequent queries are part of the same session if they have at least one common term
  - high precision, low recall strategy

Heuristics for Recovering Search Sessions

- Rewrite classes: subsequent queries are part of the same session if they follow common reformulation patterns
  - add terms, delete terms, replace terms
    - Q1: “dog coughing after being boarded”
    - Q2: “dog kennel cough”
    - Q3: “kennel cough remedies”
  - Q1-Q2 and Q2-Q3 follow common reformulation patterns
  - Q1 and Q3 have no terms in common, but are still considered part of the same session.

Exploiting Search Sessions: Sneak Preview



What about clicks?

- Explicit relevance feedback: asking the user whether a result is relevant/non-relevant to a query
- Implicit relevance feedback: predicting relevance based on user interactions

- People don't like to provide explicit feedback
- Can we use clicks to predict relevance?
  - non-obtrusive
  - inexpensive
  - lots of data

Implicit Relevance Feedback

- Question: can we use clicks to predict relevance?
- Answering this question requires understanding how users behave
- Are all clicks equally predictive of relevance?
- Are there other "forces" (other than relevance) that motivate us to click on certain results?
- What does click position tell us about where the user looked but didn't click?
- Applications: on-line learning, session-based retrieval (remember the TREC Session Track)

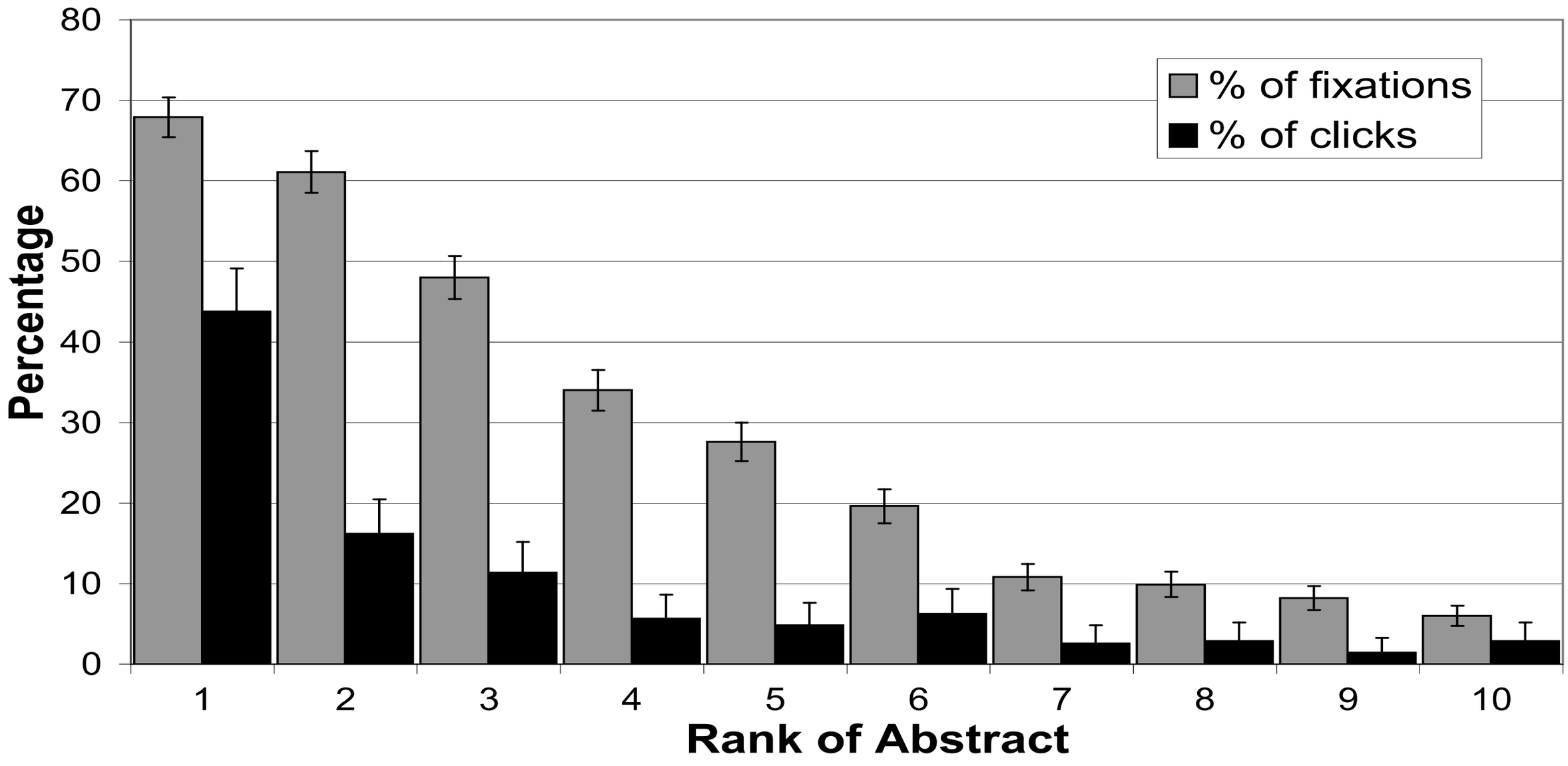
Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- First Study (Phase I)
  - 34 subjects (all Cornell undergrads)
  - 10 search tasks (5 navigational + 5 informational)
  - top-10 Google results
  - Eye tracking + click-logging
  - Fixation: spatially stable gaze lasting approximately 0.2-0.3 seconds

Implicit Relevance Feedback

(Joachims *et al.*, 2005)

- Which results do users view and click?



- % of searches where user fixated/clicked a result in rank *r*

Eye Tracking
(Joachims *et al.*, 2005)

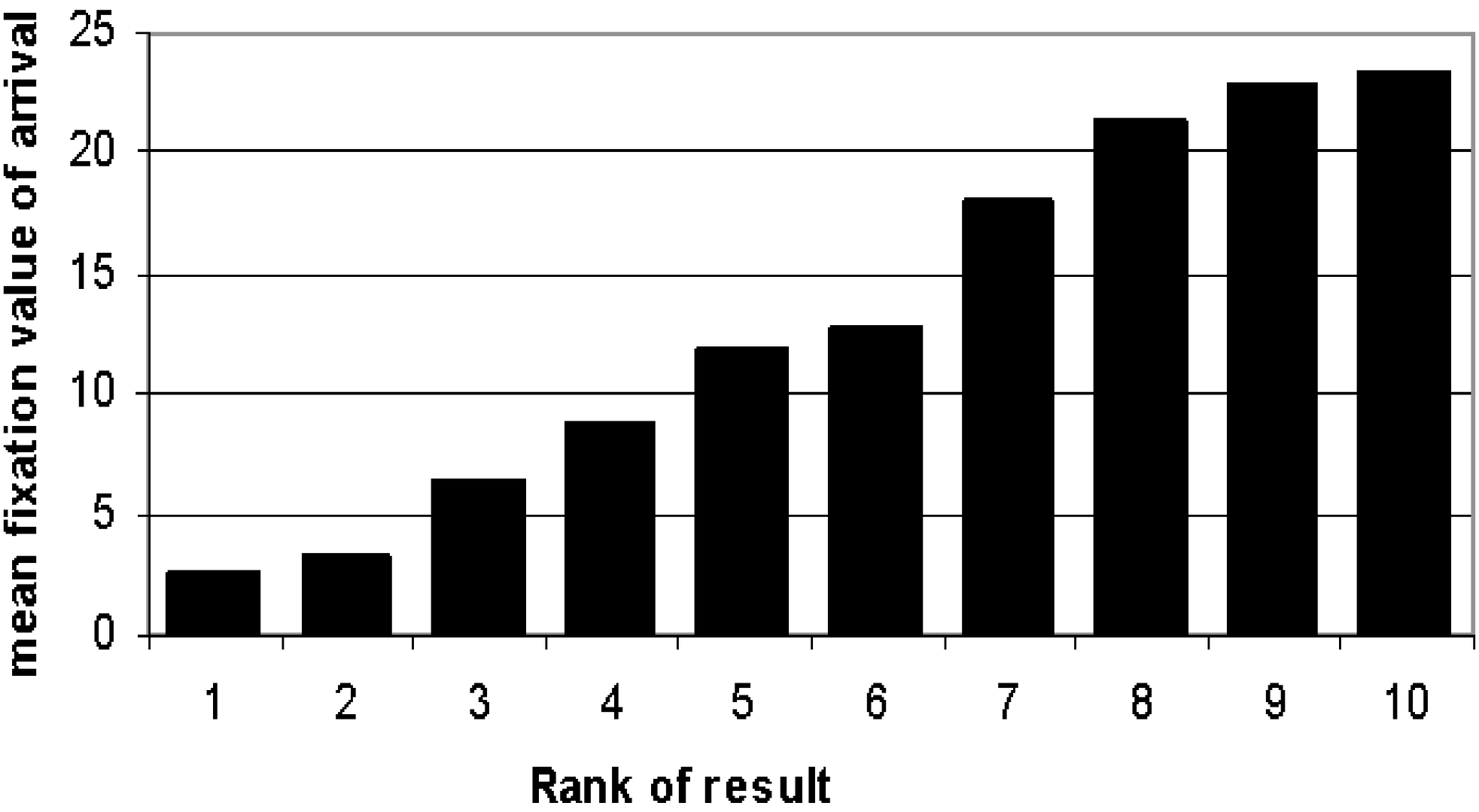
- Which results do users view?
- Most people view the first two results (almost equally)
- Fewer than half view the third result!
- Only about 10% scroll down to view results below the fold!
- Views below the fold are fairly evenly distributed. Any ideas why?

Clicks
(Joachims *et al.*, 2005)

- Which results do users click?
- While the top-two results are viewed almost equally, the first result is clicked a lot more than the second
  - Why? Because the first result is better? Because people trust it more?
- Clicks below rank 3 are fairly evenly distributed

Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- Users scan results from top to bottom



Implicit Relevance Feedback

(Joachims *et al.*, 2005)

- Which results do users evaluate before clicking?

Viewed Rank	Clicked Rank					
	1	2	3	4	5	6
1	90.6%	76.2%	73.9%	60.0%	54.5%	45.5%
2	56.8%	90.5%	82.6%	53.3%	63.6%	54.5%
3	30.2%	47.6%	95.7%	80.0%	81.8%	45.5%
4	17.3%	19.0%	47.8%	93.3%	63.6%	45.5%
5	8.6%	14.3%	21.7%	53.3%	100.0%	72.7%
6	4.3%	4.8%	8.7%	33.3%	18.2%	81.8%

Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- Users tend to look close to where they click.
- They view higher-ranks before clicking on a result
- They do so less for lower-ranked clicks.
- They also look at the one ranked immediately below the clicked result (if there is one)

Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- Second Study (Phase II)
  - 22 subjects (all Cornell undergrads)
  - same 10 tasks (5 navigational + 5 informational)
  - top-10 Google results (all results judged by assessors)
  - 3 conditions
    - normal: Google results 1-10
    - swapped: Google results 1 and 2 swapped
    - reversed: results 1-10 reversed

Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- Are clicks influenced by relevance (or just rank)?
- Relevance matters
- In the “reversed” condition (Google results 1-10 reversed), lower-ranked results were clicked more often than expected

Implicit Relevance Feedback
(Joachims *et al.*, 2005)

- So, a click = a relevance judgement?
- Not quite
- Users click on rank 1 more than rank 2 even when rank 2 is more relevant (Trust Bias!)

- So, if there's a bias in favour of the top results, how can we use clicks to predict relevance?
- It's difficult to use clicks to predict absolute relevance
- Clicks can be used, however, to predict pairwise preferences!

Implicit Relevance Feedback

(Joachims *et al.*, 2005)

Rank	1	2	3	4	5	6	7	8	9	10
Click	✓		✓				✓	✓		✓

- Click > Skip Above: ???
- Last Click > Skip Above: ???
- Click > Earlier Click: ???
- Click > Skip Previous: ???
- Click > No Click Next: ???

Implicit Relevance Feedback

(Joachims *et al.*, 2005)

Rank	1	2	3	4	5	6	7	8	9	10
Click	✓		✓				✓	✓		✓

- Click > Skip Above: (3>2), (7>2), (7>4), (7>5), (7>6), (8>2), (8>4), (8>5), (8>6), (10>2), (10>4), (10>5), (10>6), (10>9)
- Last Click > Skip Above: (10>2), (10>4), (10>5), (10>6), (10>9)
- Click > Earlier Click: (3>1), (7>1), (7>3), (8>1),(8>3), (8>7), (10>1), (10>3), (10>7), (10>8)
- Click > Skip Previous: (3>2), (7>6), (10>9)
- Click > No Click Next: (1>2), (3>4), (8>9)

Implicit Relevance Feedback

(Joachims *et al.*, 2005)

Explicit Feedback	Abstracts					Pages
Data	Phase I	Phase II				Phase II
Strategy	"normal"	"normal"	"swapped"	"reversed"	all	all
Inter-Judge Agreement	89.5	N/A	N/A	N/A	82.5	86.4
Click > Skip Above	80.8 ± 3.6	88.0 ± 9.5	79.6 ± 8.9	83.0 ± 6.7	83.1 ± 4.4	78.2 ± 5.6
Last Click > Skip Above	83.1 ± 3.8	89.7 ± 9.8	77.9 ± 9.9	84.6 ± 6.9	83.8 ± 4.6	80.9 ± 5.1
Click > Earlier Click	67.2 ± 12.3	75.0 ± 25.8	36.8 ± 22.9	28.6 ± 27.5	46.9 ±13.9	64.3 ±15.4
Click > Skip Previous	82.3 ± 7.3	88.9 ± 24.1	80.0 ± 18.0	79.5 ± 15.4	81.6 ± 9.5	80.7 ± 9.6
Click > No Click Next	84.1 ± 4.9	75.6 ± 14.5	66.7 ± 13.1	70.0 ± 15.7	70.4 ± 8.0	67.4 ± 8.2

- % agreement with pairwise preferences derived from relevance judgements from assessors
- Best strategy: a clicked result is more relevant than all higher ranked results that were skipped (not clicked)
  - produces lots of preferences that also happen to agree with explicit judgements

Conclusions and Implications
(Joachims *et al.*, 2005)

- Users' clicking decisions are influenced by relevance
- But, they are also biased in favour of the top results (the ones noticed and the ones trusted)
- Clicks should not be used to derive absolute relevance judgements
- However, they can be used to derive pairwise preference judgements!
- How could we use pairwise preference judgements (derived from clicks) to improve a search engine?

Reading

- Section 9.1 in Manning, Raghavan & Schuetze, “Modern Information Retrieval”, Cambridge University Press, 2009.
- Joachims et al. “Accurately interpreting clickthrough data as implicit feedback”, In Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 154-161, 2005.

Acknowledgements

- Based on slides prepared by Jaime Arguello