Search Log Analysis

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

Methods for IR Experimentation and Evaluation

- Test-collection (batch) evaluation
- User studies
- Search log analysis

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?

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

What is a Search-Log?

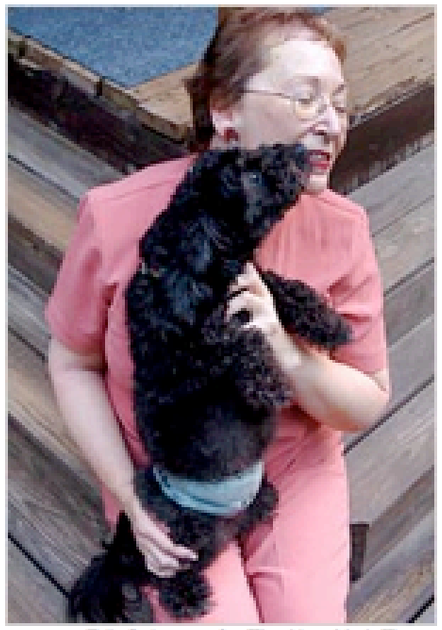
A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

Descendants
COMING SOON

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.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Multimedia

Graphic: What Revealing Search Data Reveals

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.

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

Search-Logs and Privacy

- It's 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?

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
1024071	locating serial number on rolex	2006-05-30 21:51:34	1	http://www.qualitytyme.net

(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
1024071	locating serial number on rolex	2006-05-30 21:51:34	1	http://www.qualitytyme.net

what kinds of things could we do with this?

(AOL query-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