


# **Access to Query Logs – An Academic Researcher's Point of View**

Judit Bar-Ilan, Bar-Ilan University

# The World Wide Web

- Indispensable source for information and research
- Search engines – major tools for locating information
- Need to understand how users interact with search engines





# How can we learn about users' search behavior?

- Surveys
  - Reports based on logging users activities
  - Qualitative user studies
  - Query log analyses
- 
- Each method has pros and cons
  - Need to triangulate using different methods

# Need for query logs

- Query logs are needed for research on Web search behavior
- To this date, there are only a few published studies



1039375	how to snarl	2005-10-05	1
1039375	snarling	2005-10-08	3
1039375	look mean snarling	2005-10-08	3
1039375	retirement	2005-10-09	1
1039375	sweater vests handsome	2005-11-15	3
1039375	can clothing inspire success?	2005-11-15	5
1039375	blue sweater vests	2005-11-15	1
1039375	maize sweater vests	2005-11-15	1
1039375	cheap sweater vests	2005-11-15	1
1039375	hats like jack arute wears	2005-11-15	1
1039375	jack arute fedoras	2005-11-15	1
1039375	fun with wicker and rattan	2005-11-17	2
1039375	has aj hawk hurt any QBs?	2005-11-18	1
1039375	drew henson eligibility	2005-11-19	1
1039375	todd harris abc dumb questions	2005-11-19	3
1039375	press conference snarl	2005-11-19	2
1039375	1-4	2005-11-20	1
1039375	one and four	2005-11-20	3
1039375	retirement	2005-11-20	1
1039375	evaluating assistants	2006-01-15	3
1039375	is it me or my assistant	2006-01-15	2
1039375	how should I fire someone	2006-01-15	1
1039375	online job postings assistant	2006-02-13	5
1039375	referee blind	2006-05-12	1
1039375	official visual disability	2006-05-12	1
1039375	woodson booster money 1997	2006-06-06	9
1039375	nipple clamp or nipple restraint	2006-08-01	1
1039375	retirement	2006-08-30	1
1039375	game day espn desmond's picks	2006-09-02	1
1039375	is john saunders biased?	2006-09-02	1
1039375	john saunders publicist contact	2006-09-02	1
1039375	why is the sky blue	2006-09-26	3
1039375	retirement	2006-09-26	1

# Large-scale query log analyses

- Only a few and rather “old”
- Silverstein et al. (1999)
- A series of studies by Spink and Jansen (1999-2002 Excite, AltaVista, AlltheWeb; 2004 Vivisimo)
- Recent studies (Beitzel et al., 2007 and Pass et al., 2006) based on AOL data
- Last year Microsoft released a query log (15 million queries) to a limited number of researchers (RFP awardees)
  - [http://research.microsoft.com/ur/us/fundingopps/RFPs/Search\\_2006\\_RFP\\_Awards.aspx](http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP_Awards.aspx)

# Ethical issues

- Descriptive statistics provide aggregate results, e.g. average number of query terms or number of result pages viewed
  - No privacy issues (almost)
- Data mining
  - May be useful to “advance science”
  - May interfere with users’ privacy
- Search engines have privacy policies
- What do search engines do internally??

# AOL logs

- Release of “anonymized” query logs can have unforeseen results
  - Tracking down “anonymized” users based on their queries
  - Frightened/worried users
    - May try to use devices/programs to disguise their actual information needs
    - TrackMeNot (Howe & Nissenbaum, <http://mrl.nyu.edu/~dhowe/trackmenot/>)
    - Large scale and successful disguising will probably have a negative effect on search experience



# There are problems, but ...

- Still, academics want to do research based on query logs
- Suggestion
  - Setup review boards
  - Setup clear guidelines

# Interpreting the Common Rule for the Protection of Human Subjects for Behavioral and Social Science Research – NSF

## Confidentiality-Privacy Section

- Does this mean that I cannot collect or record personal identifying information?
  - Information that would not cause harm to the individual if it were known
- What's the difference between privacy and confidentiality?
  - PRIVACY refers to persons; and to their interest in controlling the access of others to themselves.
  - CONFIDENTIALITY refers to data; and to the agreements that are made about ways in which information is restricted to certain people.
- What are the major techniques for protecting confidentiality?
  - Substitute codes for personal identifiers
  - The data can be manipulated electronically, for example by encrypting data files
  - The data can also be recoded to eliminate identifiers by collapsing it into categories.

# These guidelines are not sufficient

- They will have to be adapted to take data mining into account
- Academic researchers will probably gladly comply
- What is the interest of the search engines ???
  - Better basic training for young researchers??

# A small example – rank of clicked result

