How Are We Search the World Wide Web? A Comparison of Nine Search Engine Transaction Logs

Benard J. Jansen and Amanda Spink
Presented by Xiannong Meng
For ENGR 139, Spring 2010

Search Behavior

- In this talk, we will study how people use search engines, what kind of queries were issued, how a typical user follows the returned search results, how a user may refine search queries to get more accurate results.

Outline of the Presentation

- Major findings of the research
- Result and analysis
- Methods of Study
- Conclusions

Major Results

- Major findings of Jansen and Spink's work
  - Users are viewing fewer result pages
  - Searchers on US-based search engines use more query operators than searchers on European-based search engines
  - Significant differences exist in the use of Boolean operators and result pages viewed
  - One cannot necessarily apply results from studies of one search engine to another

What Are the Internet Users Doing?

- According to a 2008 Pew Research finding*
  - About half (49%) of the internet users use search engines in a typical day (up from one-third in 2002)
  - About 60% of the internet users use email in a typical day (up from 52% in 2002)
  - The search engine users growth rate is faster than that of email users

* http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx

What Are the Internet Users Doing? (2)

- Other users of internet
  - 39% check news
  - 30% check weather
  - 29% research hobby
  - 28% surf the web for fun
  - 13% visit social network sites

Who Are the Search Engine Users?

- Education level
  - College graduate+: 66%
  - Some college: 49%
  - High school graduate or less: 32%
- Income level
  - $75,000+: 62%
  - $50,000 – 74,999: 56%
  - $30,000 – 49,999: 34%
  - <$30,000: 36%

7

Who Are the Search Engine Users? (2)

- Age group
  - 18 – 29 years: 55%
  - 30 – 49 years: 54%
  - 50 – 64 years: 40%
  - 65 years and older: 27%
- Gender
  - Men: 53%
  - Women: 45%

8

User Query Trend

- What kind of questions a typical search engine user is asking? A recent study* shows the following changes from 1997 to 2005. (This can only be viewed as an estimate as the studies are done for different search engines.)
  - Commerce, travel, economics, and employment related queries rise from 13 percent to 30 percent
  - Entertainment and recreational related queries drop from 20 percent to 7 percent
  - Sex and pornography relate queries drop from 17 percent to 4 percent

* Spink, A. & Jansen, B.J. (2008). Trends in Searching for Commerce Related Information on Web Search Engines. *Journal of electronic commerce research*. 9(2), 154-161.

9

Statistics about Queries and Terms

- The data used in the study points out the followings.
  - The number of sessions recorded rise from 211 thousands to 535 thousands
  - The number of queries recorded rise from 1 million to 1.5 millions
  - The number of different terms used rise from 1.3 millions to 4.3 millions

10

More Details About Query Terms

- Though the search trend moves towards more e-commerce related queries, most frequently queried terms are not e-commerce related.
  - Top 20 query terms in 1997: and, of, sex, free, the, nude, pictures, in, university, pics, chat, for, adult, women, new, xxxx, girls, music, porn, to
  - Top 20 query terms in 2005: of, then, in, and, free, for, a, to, girls, sex, on, how, nude, lyrics, music, new, pictures, mp3, what, is

11

Query Operators

- US-based search engines such as AltaVista and Excite saw more query operator usage (about 20 percent of the users use them) while European-based search engines such as AllTheWeb and Fireball saw varied usage of query operator between 2 to 10 percent.

12

Other Findings

- The great majority of web queries for e-commerce and other purposes are short, infrequently modified and simple in structure.
- Few queries include advanced search features.
- Many queries contain spelling errors.
- Users general view very few results, rarely go beyond first or send page of results.

13

Method of Study

- The authors summarized nine different web query studies between 1997 and 2005, all of which involved both authors
- These nine studies used web usage logs from various search engines including *Excite, Fireball, AlltheWeb, AltaVista,* after anonymizing the data.
- A session is recorded as one sequence of continues interactions with the search engine
- Collectively, nearly 300 millions of web search sessions, and over a billion of queries are represented in the studies

14

Method of Study (2)

- Terms are categorized into 11 classes
  1. People, places, or things
  2. Commerce, travel, employment, or economy
  3. Computers or internet
  4. Health or science
  5. Education or humanities
  6. Entertainment or recreation
  7. Sex and pornography
  8. Society, culture, ethnicity, or religion
  9. Government or legal
  10. Performing or fine arts
  11. Non-english or unknown

15

Term Classification

- The *commerce* category includes product and company names such as “coca cola”, “Walmart”, and “Rolex”.
- The *travel* category includes, “railway trains in Canada”, “beach vacation Bahamas” and “cheap flights”.
- The *employment* category includes, “jobs in Miami”, “how to join the army” and “child care positions”.
- The *economic* category includes finance queries, such as “financial planning”, “government bonds’ and “stock market falls”.

16

Session and Query Length

- In a separate but related study*, the authors have the following findings
  - Over 50 percent of the sessions are one-query only.
  - The percentage of one-query session dropped from 60 percent in 1999 to 55 percent in 2002.
  - The percentage of one-term query is somewhere between 20 to 30 percent.
  - The percentage of users viewing one-page results increases from 29 percent in 1997 to 73 percent in 2002 for U.S. population.

* Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1), 248-263.

17

Number of Terms, Queries, and Users

- In yet another study*, the authors find
  - The number of terms per query changes little: 1997, 2.4; 1999, 2.4; 2001, 2.6;
  - The number of queries per user: 1997, 2.5; 1999, 1.9; 2001, 2.3;
  - The number of pages viewed per query: 1997,1.7; 1999, 1.6; 2001, 1.7;
  - Percentage of users who modify queries: 1997, 52.0; 1999, 39.6; 2001, 44.6;
  - Percentage of users who use top-100 frequent query terms: 1997, 17.9; 1999, 19.3, 2001 22.0.

* Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, 35(3), 107 - 109.

18

How Data Were Collected?

- The data used in the studies were collected from search engine logs of five major search engines in nine different studies between 1997 and 2002.
- A search engine essentially runs as server program. It can log user query and host information by using calls such as getPeerInformation in Java or in C/C++.

19

Internet and IP Addresses

- There are two types of IP address
  - An IPv4 address has 4 bytes (4 numbers if you wish) in the form of 134.82.56.104, each segment is in the range of 0-127. This is the most popular in use.
  - An IPv6 has 16 bytes, in a similar form to that of IPv4, only much longer. It has about one percent of the total internet address.
- An IP address is assigned by the network authority ICANN (ICANN - Internet Corporation for Assigned Names and Numbers)

20

IP Address and Computer Name

- It would be very difficult for humans to use IP address directly (do you want to try 64.233.169.103 instead of www.google.com?)
- Solution? Name and address conversion by specialized computers (or more accurately software).
- A software system called DNS (Domain Name System) is responsible for this translation.

21

An Example of IP in Action

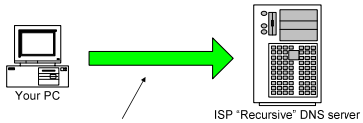
- Retrieved and revised from: http://www.aptld.org/file/APTLD_DNS.ppt

22

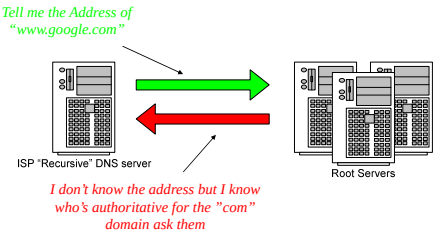
Accessing a Web Page

- You type http://www.google.com into your web browser and hit enter.
- What happens now?

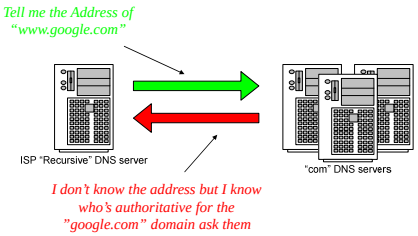
Step 1: Your PC sends a resolution request to its configured DNS Server, typically at your ISP.



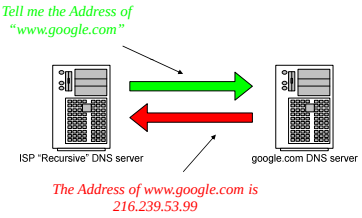
Step 2: Your ISPs recursive name server starts by asking one of the root servers predefined in its “hints” file.



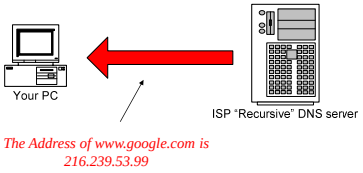
Step 3: Your ISPs recursive name server then asks one of the “com” name servers as directed.



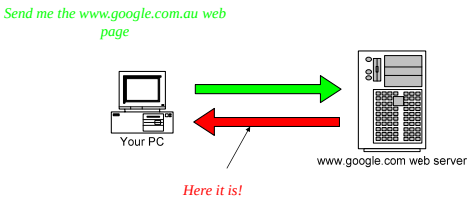
Step 4: Your ISPs recursive name server then asks one of the “google.com” name servers as directed.



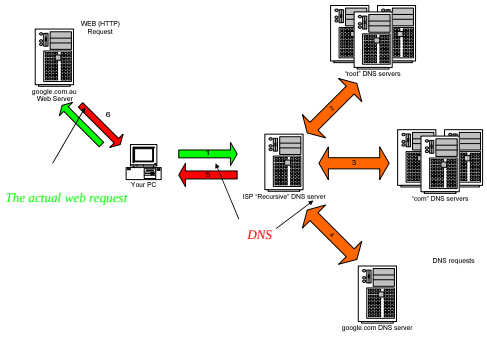
Step 5: ISP DNS server then sends the answer back to your PC. The DNS server will “remember” the answer for a period of time.



Step 6: Your PC can then make the actual HTTP request to the web server.



IP and DNS Summary

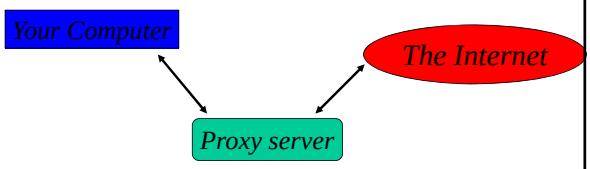


What That Has Anything To Do with Web Logs?

- When the two computers start to communicate with each other, they have to use each other's IP addresses to send information back and forth.
- Web log is what the host computer records the information about the visitors.
- Current legislation allows web server to log the information.
- Not to be logged? Use proxy servers.

31

Proxy Servers



32

Conclusions

- Users are viewing fewer result pages
- Searchers on US-based search engines use more query operators than searchers on European-based search engines
- Sessions are relatively short for average search engine users
- One cannot necessarily apply results from studies of one search engine to another

33

References (1)

- Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248-263. Retrieved August 4, 2009 from: http://dx.doi.org/10.1016/j.ipm.2004.10.007
- Fallows, D. (August 2008). Search Engine. *Pew Research Report.* Accessed February 17, 2010. http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx
- Spink, A. & Jansen, B.J. (2008). Trends in Searching for Commerce Related Information on Web Search Engines. *Journal of electronic commerce research*. 9(2), 154-161.

34

References (2)

- Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, 35(3), 107 - 109.

35