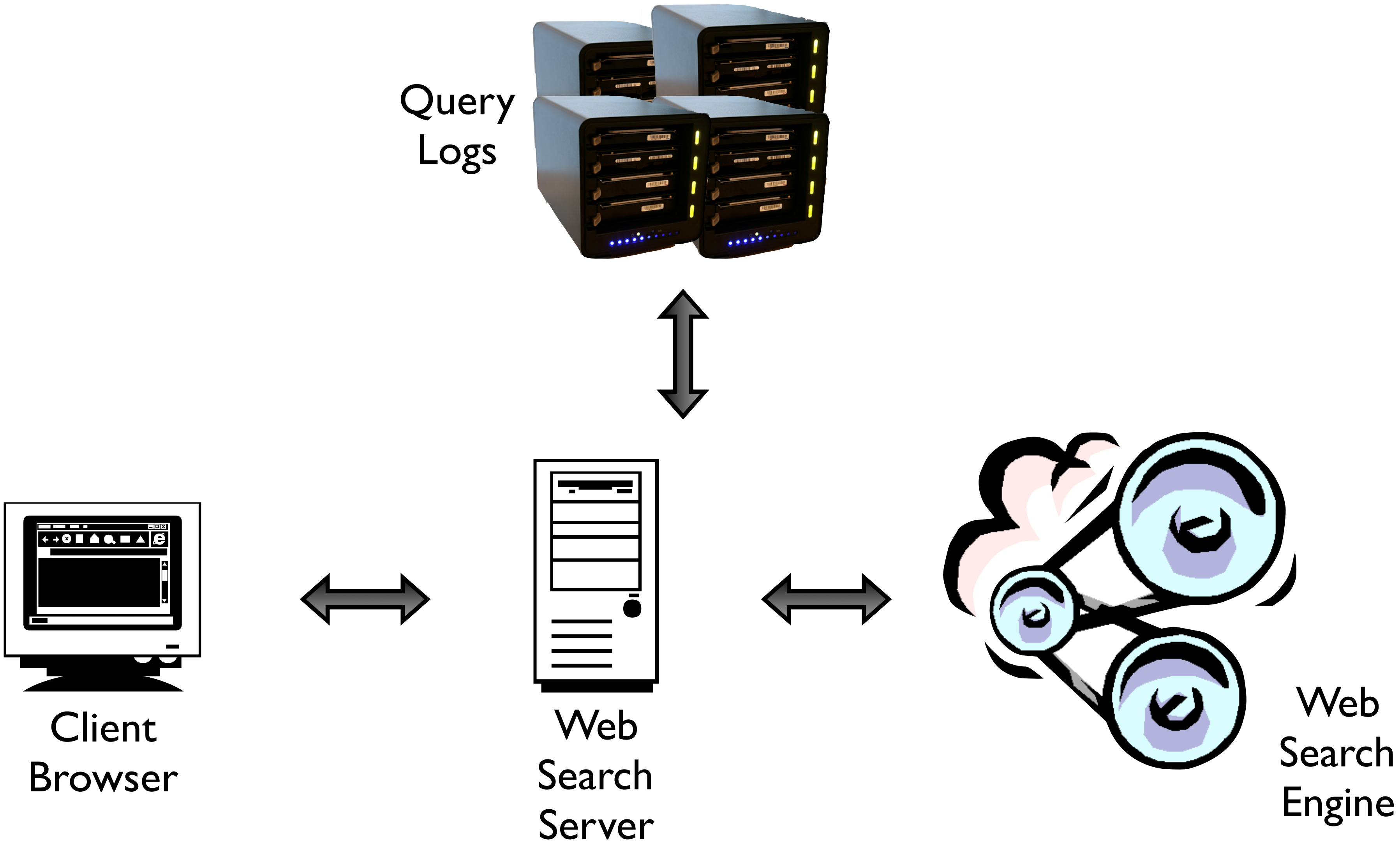
Mining di Dati Web

Lezione 4: Query Logs

Introduction



Query Logs

Client Browser

Web Search Server

Web Search Engine

What's Stored into Query Logs?

- The query
- The requested page of results
- The searcher ID (usually IP address)
- The clicked result
- The clicked URL

What’s Really Stored into Query Logs?

- Our tastes
- Our preferences
- Our desires
- Our curiosities
- ...

So... What's REALLY REALLY there?

*User #17988817 searched on '**how to kill someone**' found http://www.43things.com at 2006-04-26 20:41:06*

*User #17988817 searched on '**how to kill someone**' found http://www.everything2.com at 2006-04-26 20:41:06*

*User #17988817 searched on '**how to kill someone**' found http://quizilla.com at 2006-04-26 20:41:06*

*User #17988817 searched on '**how to kill someone**' found http://www.linkbase.org at 2006-04-26 20:41:06*

*User #17988817 searched on '**the perfect murder**' found http://aperfectmurder.warnerbros.com at 2006-04-26 20:40:16*

*User #17988817 searched on '**the perfect mudred**' at 2006-04-26 20:39:57*

*User #17988817 searched on '**the real perfect crime**' found http://www.wired.com at 2006-04-26 20:37:38*

*User #17988817 searched on '**the perfect crime**' found http://www.perfect-crime.com at 2006-04-26 20:35:58*

*User #17988817 searched on '**real serial killers**' found http://www.murraymoffatt.com at 2006-04-26 20:34:57*

*User #17988817 searched on '**real serial killers**' found http://members.firstinter.net at 2006-04-26 20:27:34*

*User #17988817 searched on '**real serial killers**' found http://www.geocities.com at 2006-04-26 20:27:34*

*User #17988817 searched on '**serial killers**' at 2006-04-26 20:26:23*

*User #17988817 searched on '**killers**' at 2006-04-26 20:23:41*

*User #17988817 searched on '**google**'*

Uhmmm!!!!

*User #2732442 searched on '**how do i know if my friend is a lesbian**' found http://www.avert.org at 2006-03-07 01:31:52*

*User #2732442 searched on '**do lesbians have loud voices**' at 2006-03-07 01:29:30*

*User #2732442 searched on '**what do you do if a girl kisses you and they are the same sex as you**' at 2006-03-07 01:27:15*

*User #2732442 searched on '**can you kiss your friends**' at 2006-03-07 01:22:55*

WTF!?!

*User #582088 searched for '****can you hear me out there i can hear you i got you i can hear you over i really feel strange i wanna wish for something new this is the scariest thing ive ever done in my life who do we think we are angels and airwaves im gonna count down till 10 52 i can****' at 2006-05-30 19:31:54*

I Think I’m Becoming Addicted

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**what is obsessive compulsive disorder**' at 2006-05-08 08:37:14*

*User #11486558 searched on '**cat wiping butt on carpet**' at 2006-05-08 08:28:18*

*User #11486558 searched on '**cat wiping butt on carpet**' at 2006-05-08 08:28:18*

*User #11486558 searched on '**cat wiping butt on carpet**' at 2006-05-08 08:28:18*

*User #11486558 searched on '**my cat keeps rubbing her butt on the carpet medical problem**' at 2006-05-08 08:27:27*

*User #11486558 searched on '**anxiety and fasciculations**' at 2006-05-08 08:18:52*

*User #11486558 searched on '**twitching fasciculations at the base of thumb**' at 2006-05-08 08:18:17*

*User #11486558 searched on '**twitching fasciculations at the base of thumb**' at 2006-05-08 08:13:15*

*User #11486558 searched on '**twitching fasciculations at the base of thumb**' at 2006-05-08 08:13:15*

*User #11486558 searched on '**twitching fasciculations at the base of thumb**' at 2006-05-08 08:13:15*

*User #11486558 searched on '**twitching fasciculations at the base of thumb**' at 2006-05-08 08:13:15*

The Last One!!

*User #6416389 searched on '**cannibals eating fat girls**' at 2006-05-11 18:48:05*

*User #6416389 searched on '**cannibals eating fat girls**' found http://www.gymmuenchenstein.ch at 2006-05-11 18:44:05*

*User #6416389 searched on '**cannibals eating fat girls**' found http://www.pervscan.com at 2006-05-11 18:40:57*

*User #6416389 searched on '**cannibals eating fat girls**' found http://www.mayhem.net at 2006-05-11 18:40:57*

*User #6416389 searched on '**pictures of women being spit roasted**' at 2006-05-11 15:06:41*

*User #6416389 searched on '**pictures of women being spit roasted**' found http://www.villagevoice.com at 2006-05-11 15:01:04*

*User #6416389 searched on '**dolcett roasted girls**' found http://www.geocities.com at 2006-05-10 15:47:03*

*User #6416389 searched on '**dolcett roasted girls**' found http://www.geocities.com at 2006-05-10 15:47:03*

*User #6416389 searched on '**dolcett roasted girls**' found http://www.sickestsites.com at 2006-05-10 15:47:03*

*User #6416389 searched on '**dolcett roasted girls**' found http://www.sickestsites.com at 2006-05-10 15:47:03*

*User #6416389 searched on '**roasting girls in a barbeque pit**' at 2006-05-10 15:46:16*

*User #6416389 searched on '**roasting girls in a barbeque pit**' at 2006-05-10 15:45:47*

*User #6416389 searched on '**muki**' found http://www.mukiskitchen.com at 2006-05-10 10:59:01*

*User #6416389 searched on '**cannibal horror stories**' found http://movies.monstersandcritics.com at 2006-05-10 10:57:30*

*User #6416389 searched on '**cannibal horror stories**' found http://www.locusmag.com at 2006-05-10 10:51:45*

*User #6416389 searched on '**cannibal horror stories**' found http://daha.best.vwh.net at 2006-05-10 10:51:45*

*User #6416389 searched on '**cooking the buttocks of women**' at 2006-05-10 10:50:55*

*User #6416389 searched on '**cooking the buttocks of women**' at 2006-05-10 10:50:21*

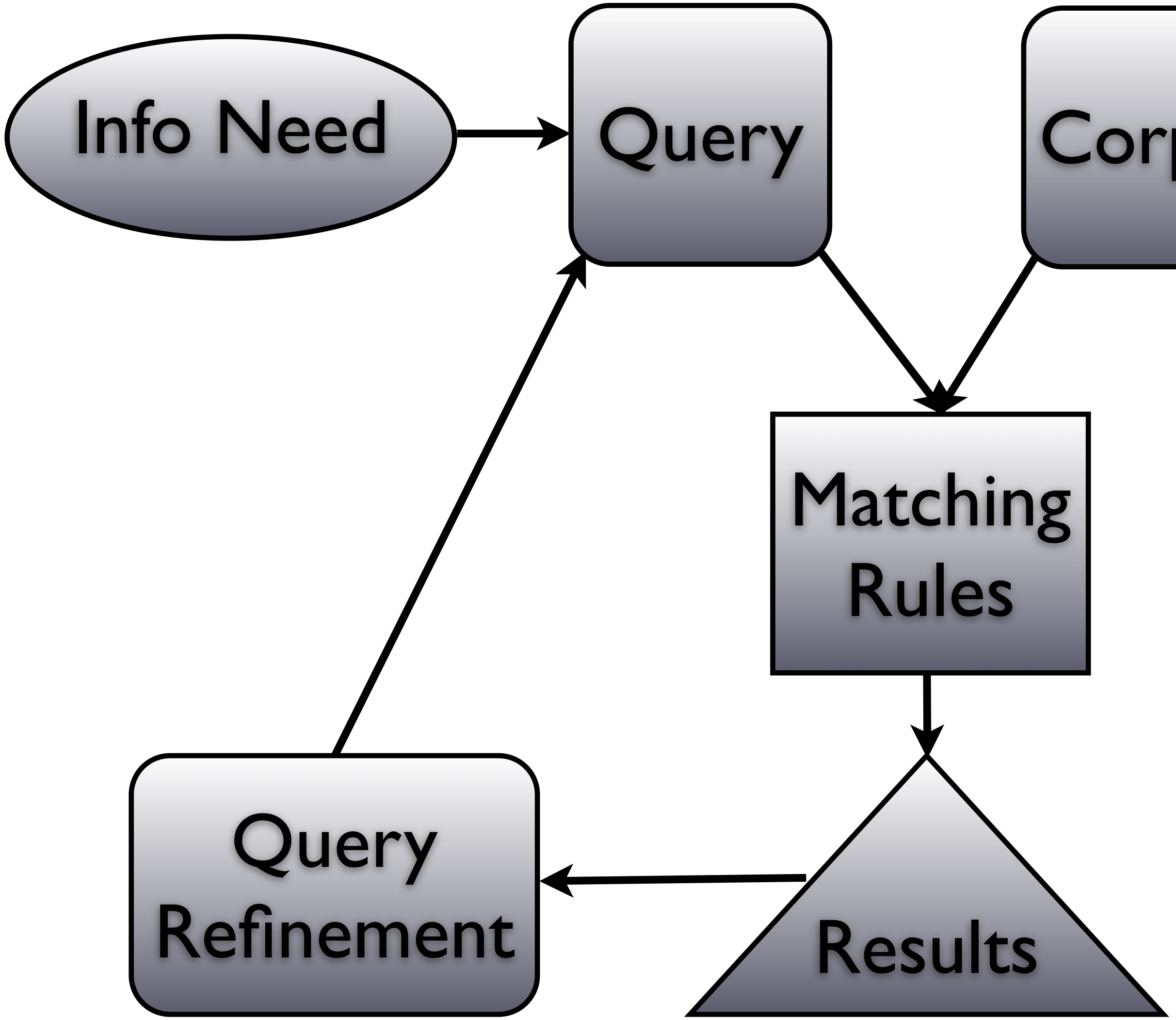
*User #6416389 searched on '**fat buttocks of girls ready for the pot**' at 2006-05-10 10:49:16*

*User #6416389 searched on '**raping and cooking girl flesh**' at 2006-05-10 10:48:32*

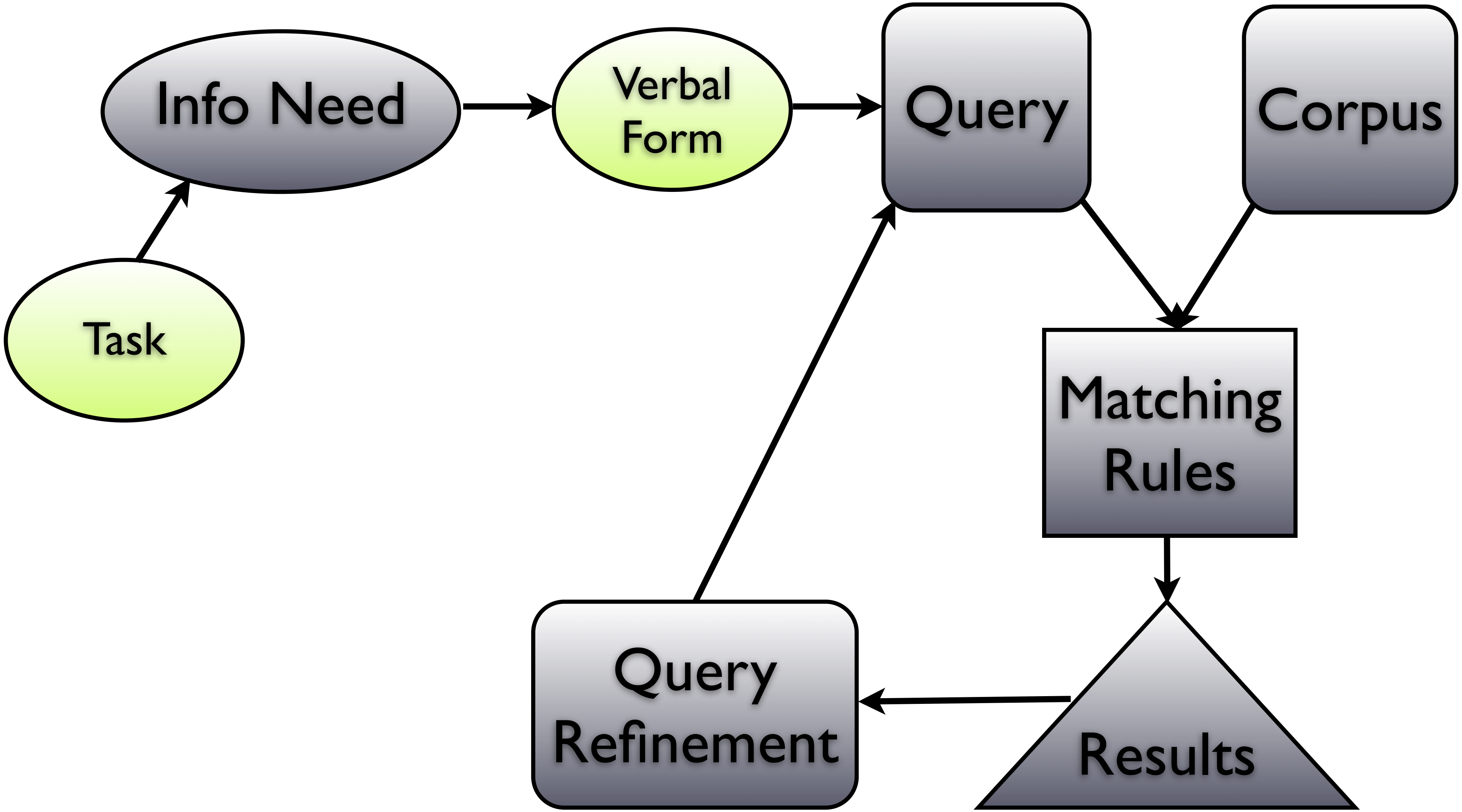
Serious Studies

- Broder, A. 2002. **A taxonomy of web search**. SIGIR Forum 36, 2 (Sep. 2002), 3-10.
- Jansen, B. J. and Spink, A. 2006. **How are we searching the world wide web?: a comparison of nine search engine transaction logs**. Inf. Process. Manage. 42, 1 (Jan. 2006), 248-263.
- Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. 2004. **Hourly analysis of a very large topically categorized web query log**. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM, New York, NY, 321-328.

Classical IR Query Formulation



Web IR Query Formulation



Verbal Form

- Means put in words what we want to search.
- Put in “simple” words like:
  - “*real serial killer*”
  - “*cat wiping butt on carpet*” instead of “*my cat keeps rubbing her butt on the carpet medical problem*”

Tasks

- Users have in mind something when search the Web
- User intent is driven by a task to accomplish

A Taxonomy of Web Searches

- Broder, A. 2002. A taxonomy of web search. SIGIR Forum 36, 2 (Sep. 2002), 3-10.
- Web queries can be classified according to their intent into 3 classes:
  - *Navigational*
  - *Informational*
  - *Transactional*

Navigational Queries

- Greyhound Bus
  - Probable target http://www.greyhound.com
- compaq
  - Probable target http://www.compaq.com
- national car rental
  - Probable target http://www.nationalcar.com
- american airlines home
  - Probable target http://www.aa.com
- Don Knuth
  - Probable target http://www-cs-faculty.stanford.edu/~knuth

Informational Queries

- Wide
    - cars
    - San Francisco
- Narrow
    - normocytic anemia
    - Scoville heat units

Transactional Queries

- Web search engines are used to look for a more specialized provider
  - find music
  - where is the yellow page DB?
  - find servers for gaming

User Survey

2. Which of the following describes best what you are trying to do?

24.53% I want to get to a specific website that I already have in mind

68.41% I want a good site on this topic, but I don't have a specific site in mind

3. Which of the following best describes why you conducted this search?

8.16% I am shopping for something to buy on the Internet

5.46% I am shopping for something to buy elsewhere than on the Internet

22.55% I want to download a file (e.g., music, images, programs, etc.)

57.19% None of these reasons

4. Which of the following describes best what you are looking for?

14.83% A site which is a collection of links to other sites regarding this topic

76.62% The best site regarding this topic

Query Classification

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

Comparison of Nine Search Engines

- Jansen, B. J. and Spink, A. 2006. How are we searching the world wide web?: a comparison of nine search engine transaction logs. Inf. Process. Manage. 42, 1 (Jan. 2006), 248-263.
- a 1997 study of Excite Web search engine
- a 1998 study of the Fireball Web search
- a 1998 study of the AltaVista Web search engine
- a 1999 study of the Excite Web search engine
- a 2000 study of the BWIE Web search service
- a 2001 study of Excite Web search engine
- a 2002 study of AlltheWeb.com
- a 2002 study of AltaVista

Characteristics of the Studies

Study no.	1	2	3	4
	Excite	Fireball	AltaVista	Excite
Region	US	European	US	US
Data collection	Tuesday 16 September 1997	1–31 July 1998	2 August–13 September 1998	Wednesday 1 December 1999
Sessions	211,063	Not reported	285,474,117	325,711
Queries	1,025,908	16,252,902	993,208,159	1,025,910
Terms	1,277,763	Not reported	Not reported	1,500,500

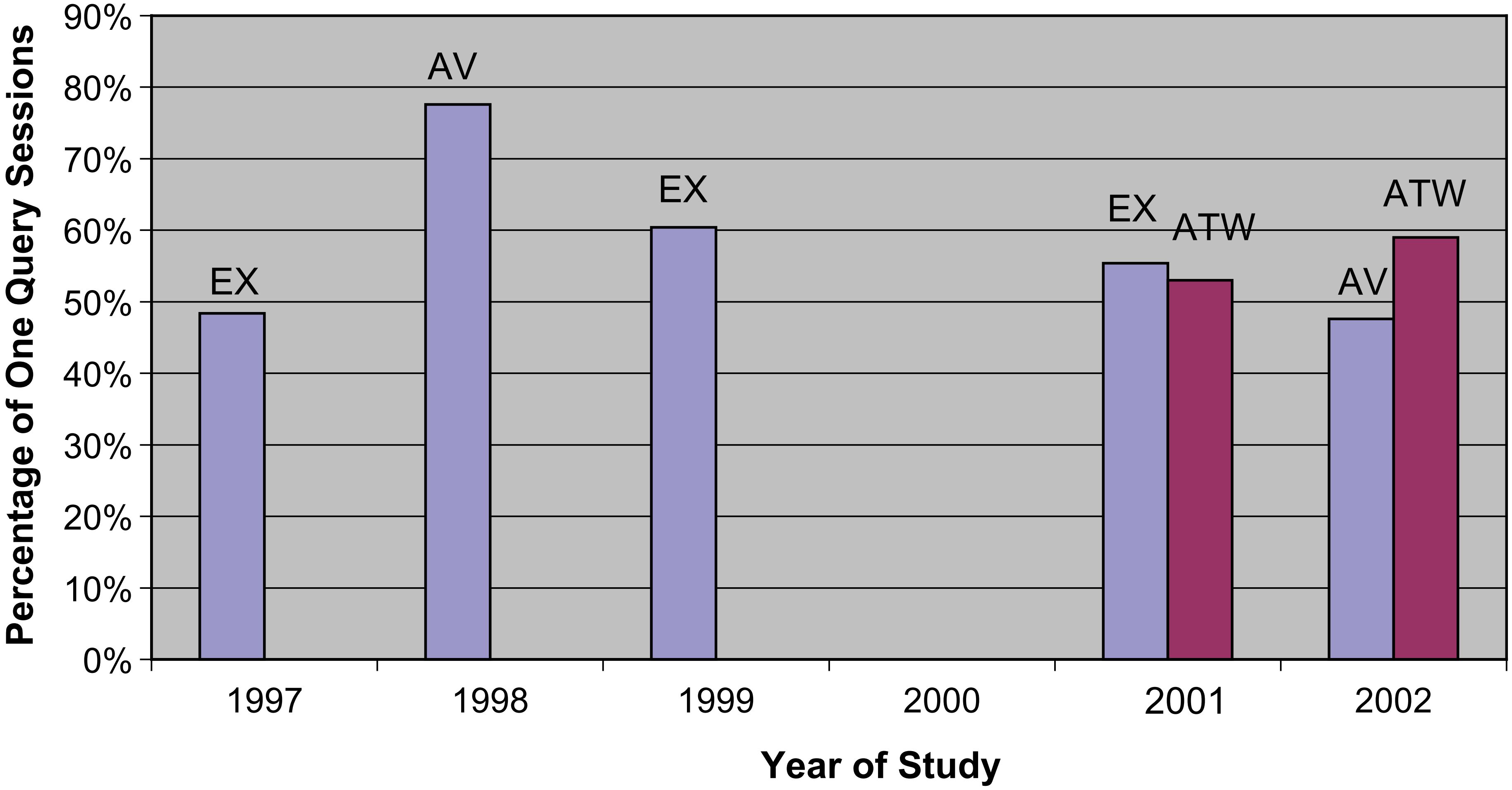
Characteristics of the Studies

Study no.	5	6	7	8	9
	BWIE	AlltheWeb.com	Excite	AlltheWeb.com	AltaVista
Region	European	European	US	European	US
Data collection	3–18 May 2000	Tuesday 6 February 2001	Monday 30 April 2001	Tuesday 28 May 2002	Sunday 8 September 2002
Sessions	83,232	153,297	262,025	345,093	369,350
Queries	71,810	451,551	1,025,910	957,303	1,073,388
Terms	116,953	1,350,619	1,538,120	2,225,141	1,073,388

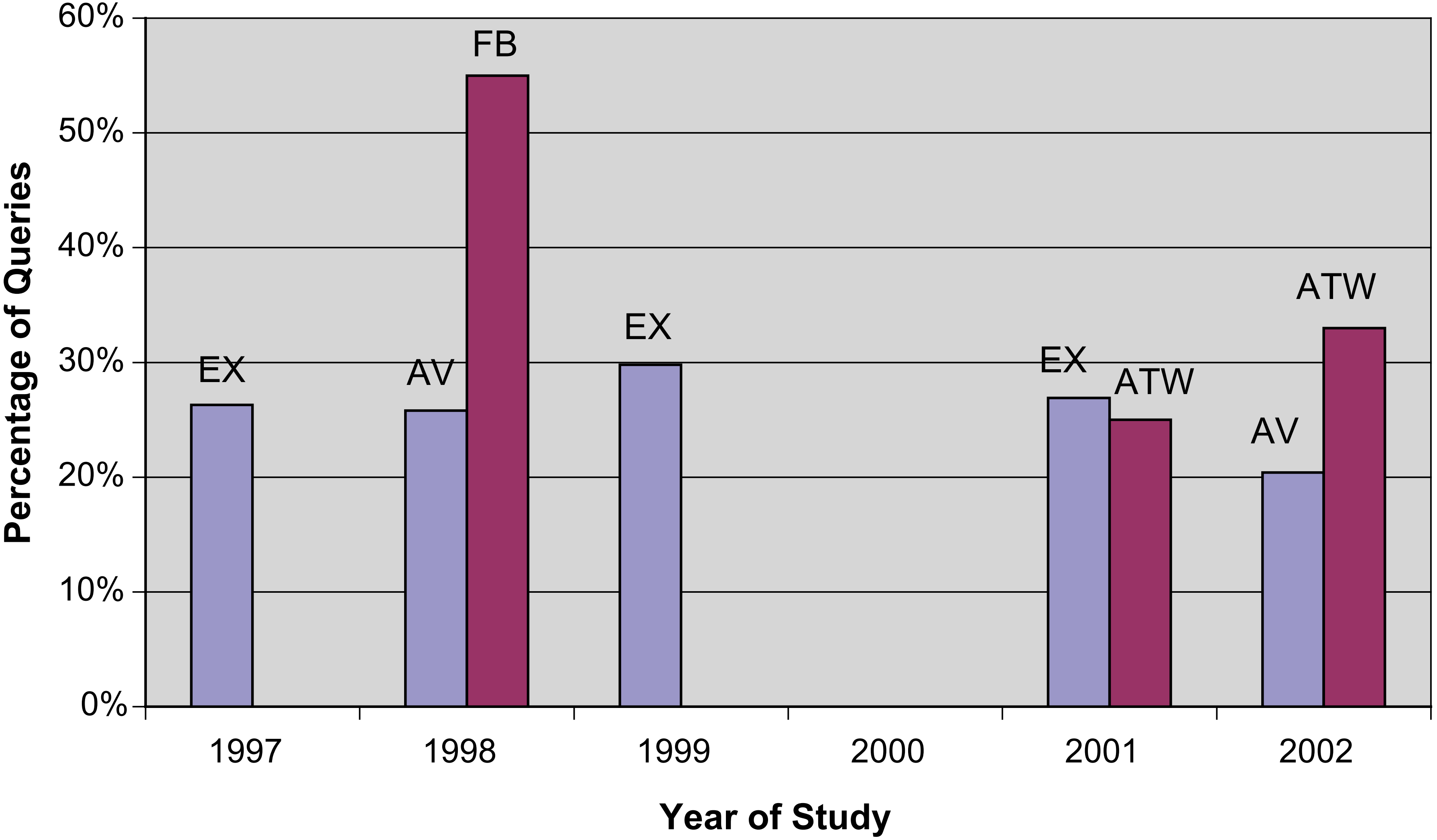
Terminology

- **Session length**
  - is the number of queries that a searcher submits in one episode with a Web search engine
- **Query length**
  - is the number of terms in a query
- **Term**
- **Operator usage**
  - whether or not the query includes boolean operators like AND, OR, MUST, APPEAR, PHRASE
- **Result page**
- **Result page viewed**
  - It usually is the clicked result

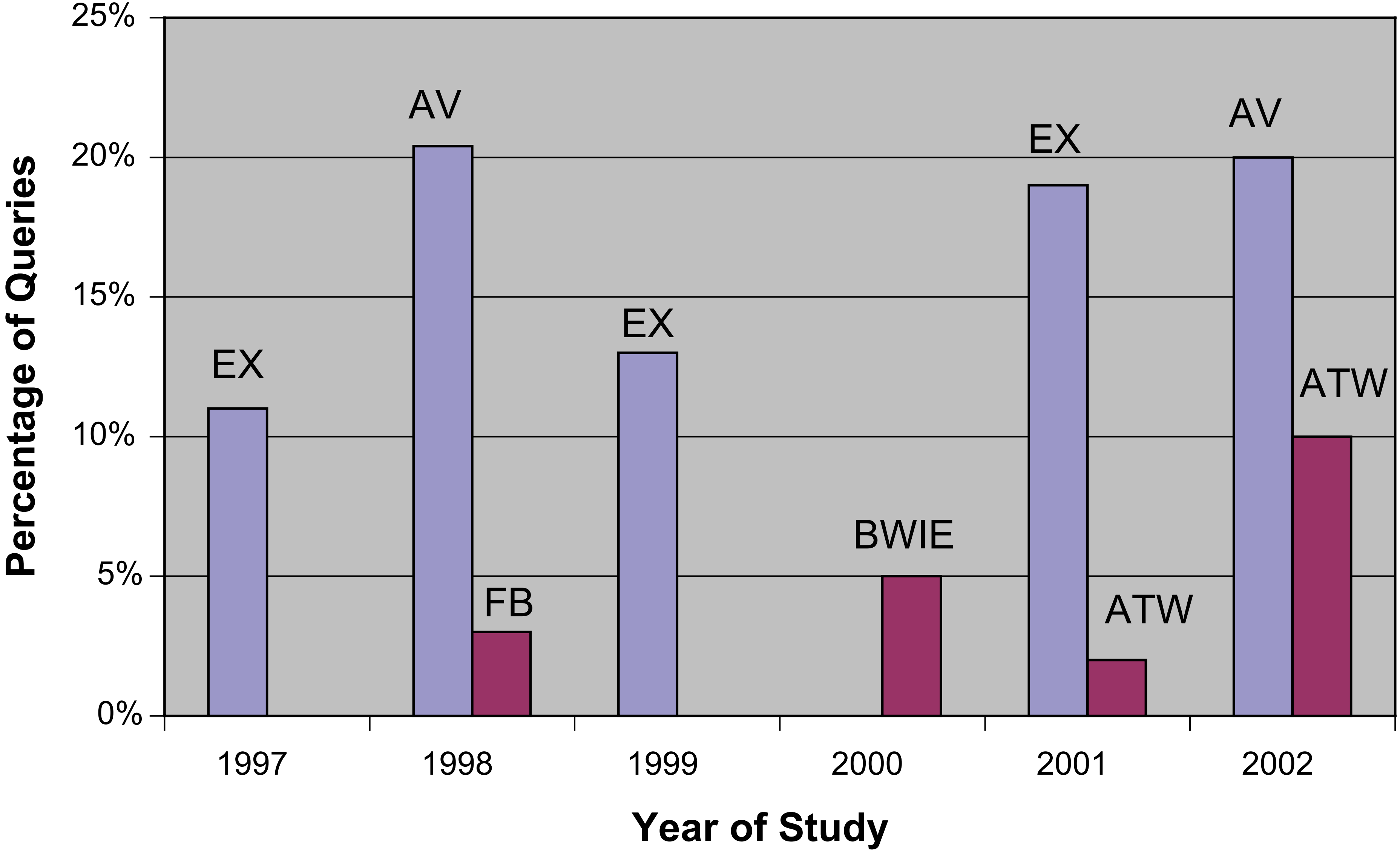
Percentage of Single Query Sessions



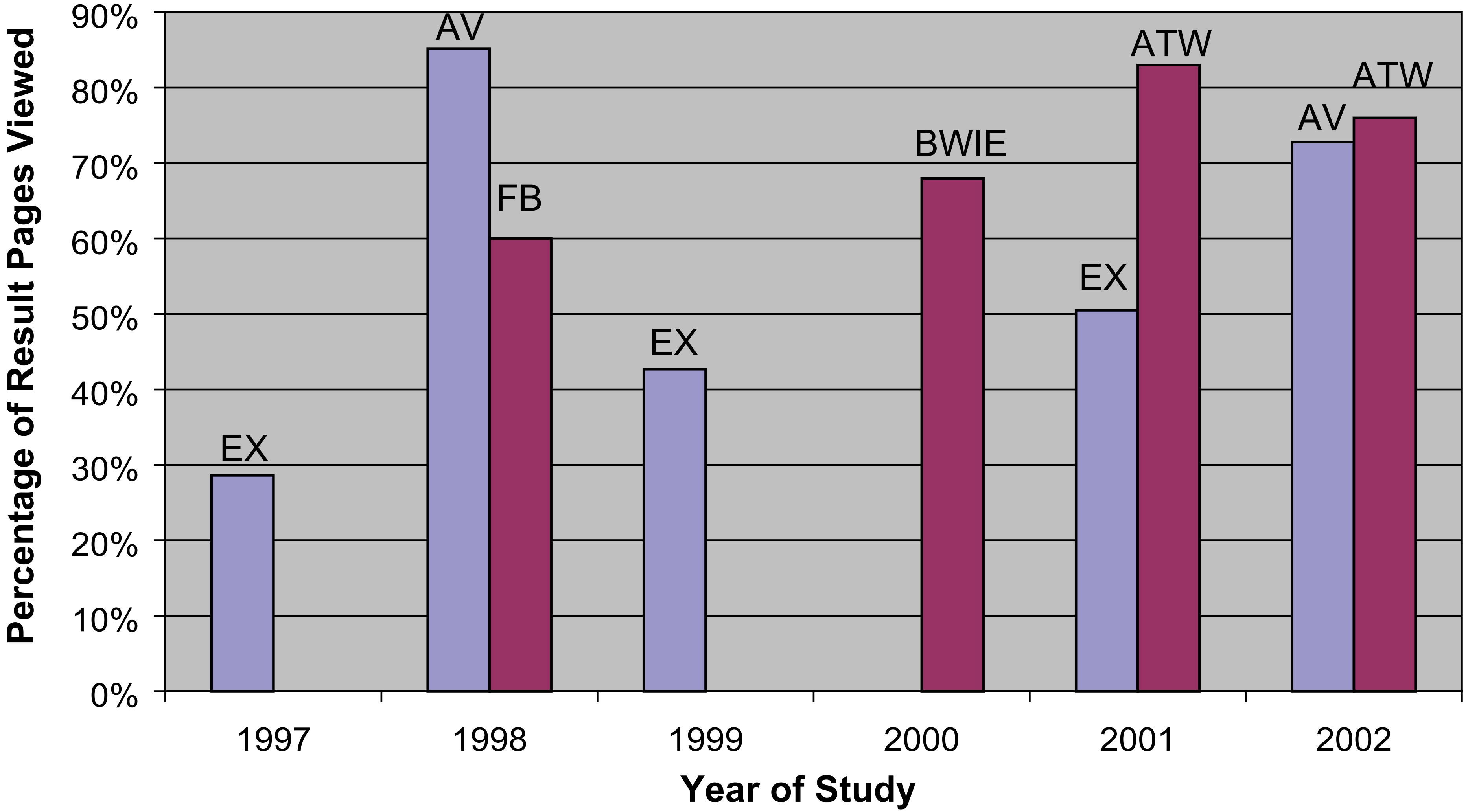
Percentage of One-Term Queries



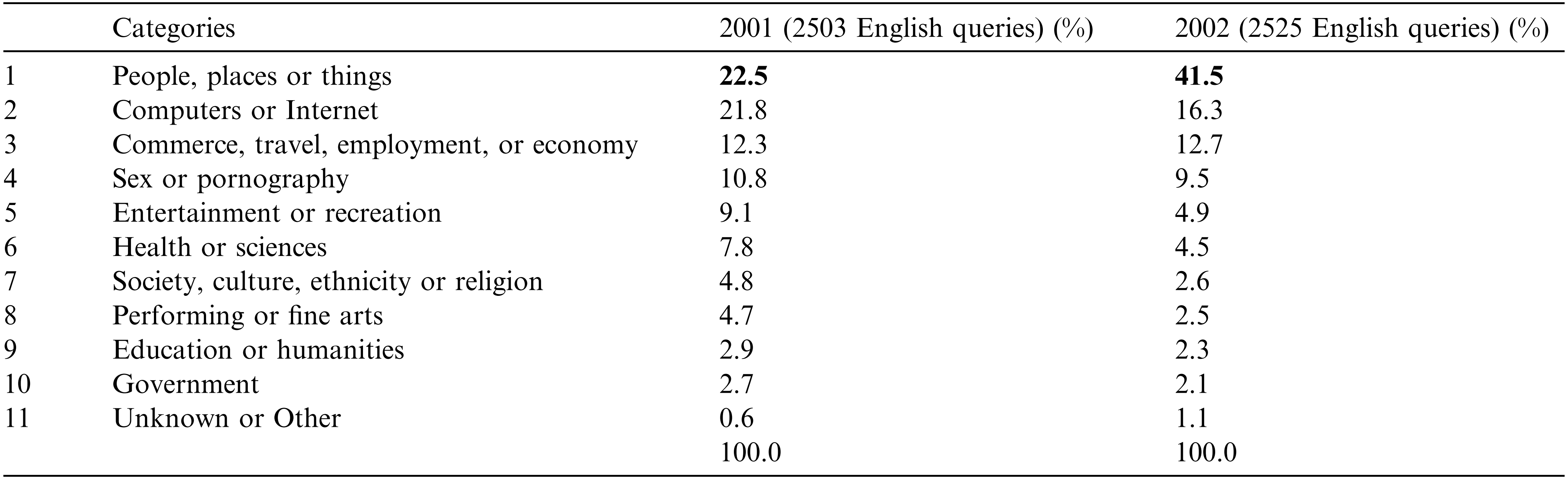
Operator Usage



Single Result Page Views



AllTheWeb Topics Queried



	Categories	2001 (2503 English queries) (%)	2002 (2525 English queries) (%)
1	People, places or things	**22.5**	**41.5**
2	Computers or Internet	21.8	16.3
3	Commerce, travel, employment, or economy	12.3	12.7
4	Sex or pornography	10.8	9.5
5	Entertainment or recreation	9.1	4.9
6	Health or sciences	7.8	4.5
7	Society, culture, ethnicity or religion	4.8	2.6
8	Performing or fine arts	4.7	2.5
9	Education or humanities	2.9	2.3
10	Government	2.7	2.1
11	Unknown or Other	0.6	1.1
		100.0	100.0

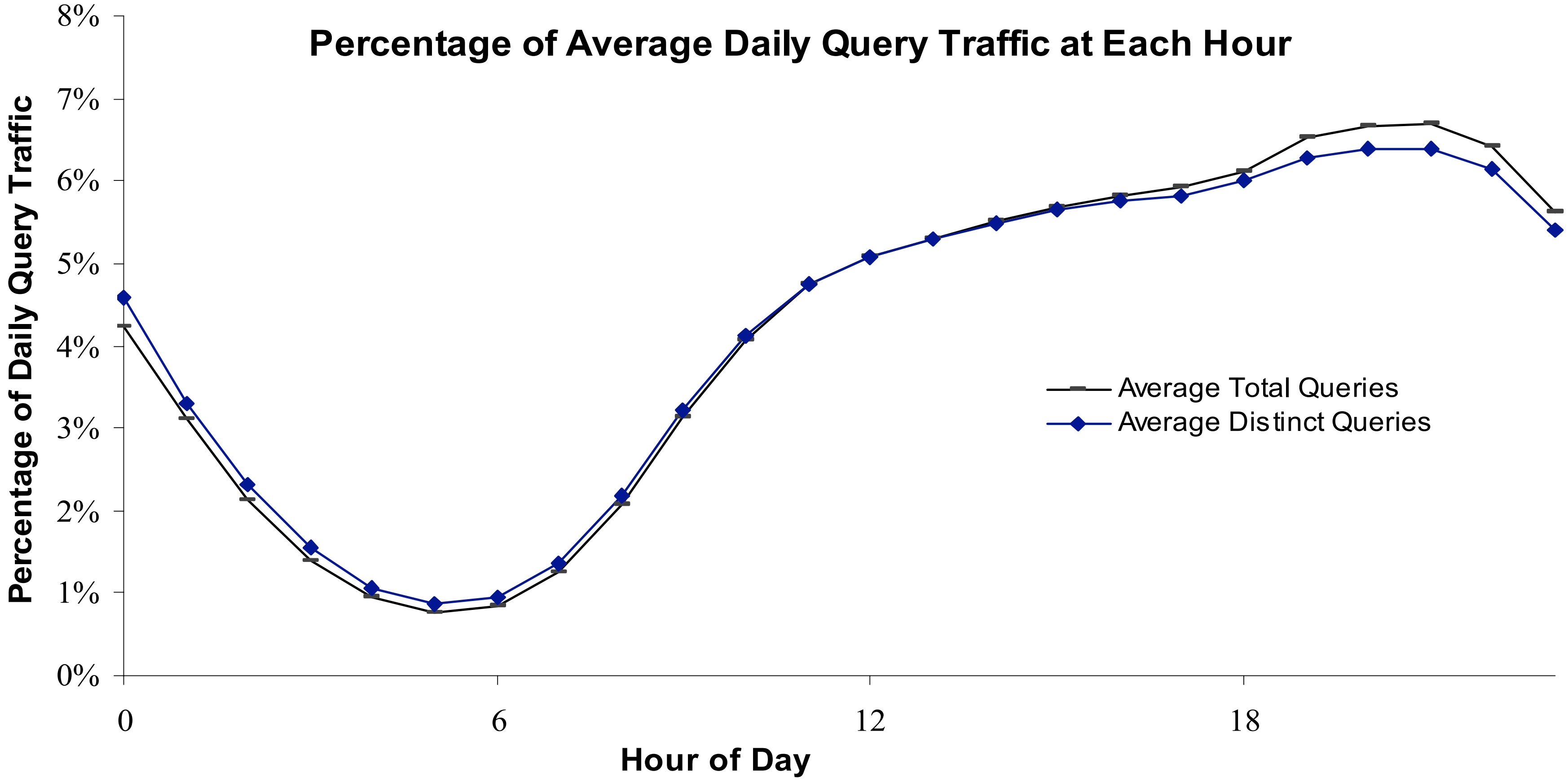
Excite AltaVista Topics

	Categories	1997 Excite (2414 queries) (%)	1999 Excite (2539 queries) (%)	2001 Excite (2453 queries) (%)	2002 AltaVista (2603 queries) (%)
1	People, places, or things	6.7	20.3	19.7	**49.3**
2	Commerce, travel, employment, or economy	13.3	**24.5**	**24.7**	12.5
3	Computers or Internet	12.5	10.9	9.7	12.4
4	Health or sciences	9.5	7.8	7.5	7.5
5	Education or humanities	5.6	5.3	4.6	5.0
6	Entertainment or recreation	**19.9**	7.5	6.7	4.6
7	Sex and pornography	16.8	7.5	8.6	3.3
8	Society, culture, ethnicity, or religion	5.7	4.2	3.9	3.1
9	Government	3.4	1.6	2.0	1.6
10	Performing or fine arts	5.4	1.1	1.2	0.7
11	Non-English or unknown	4.1	9.3	11.4	0.0
		102.9	100.0	100.0	100.0

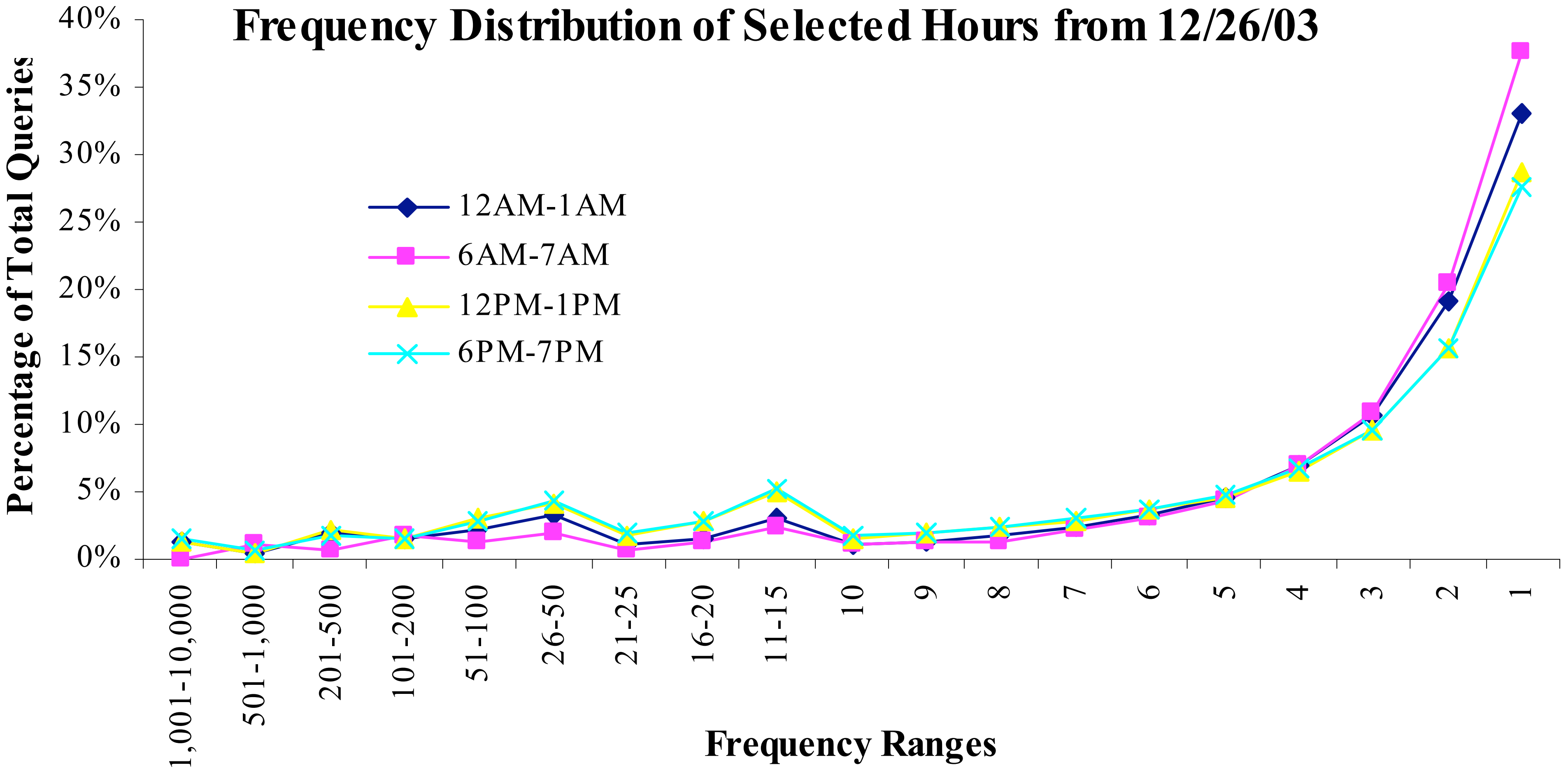
Tracking Down Query Analysis

- Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. 2004. Hourly analysis of a very large topically categorized web query log. In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM, New York, NY, 321-328.
- Analyze a query log of hundred of millions of queries
- An entire week of query traffic to AOL search service
- Analysis made on a hourly basis

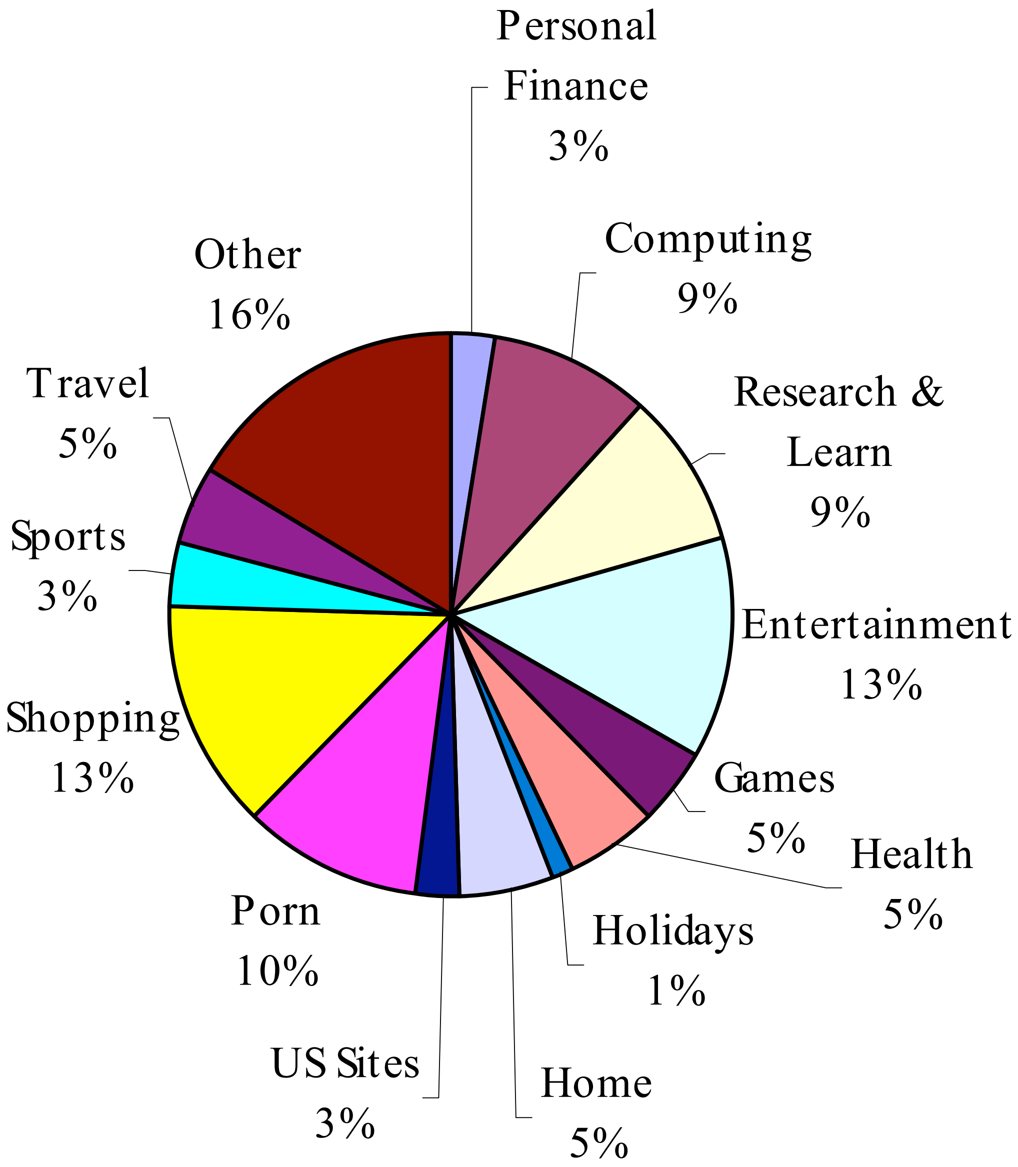
Query Distribution



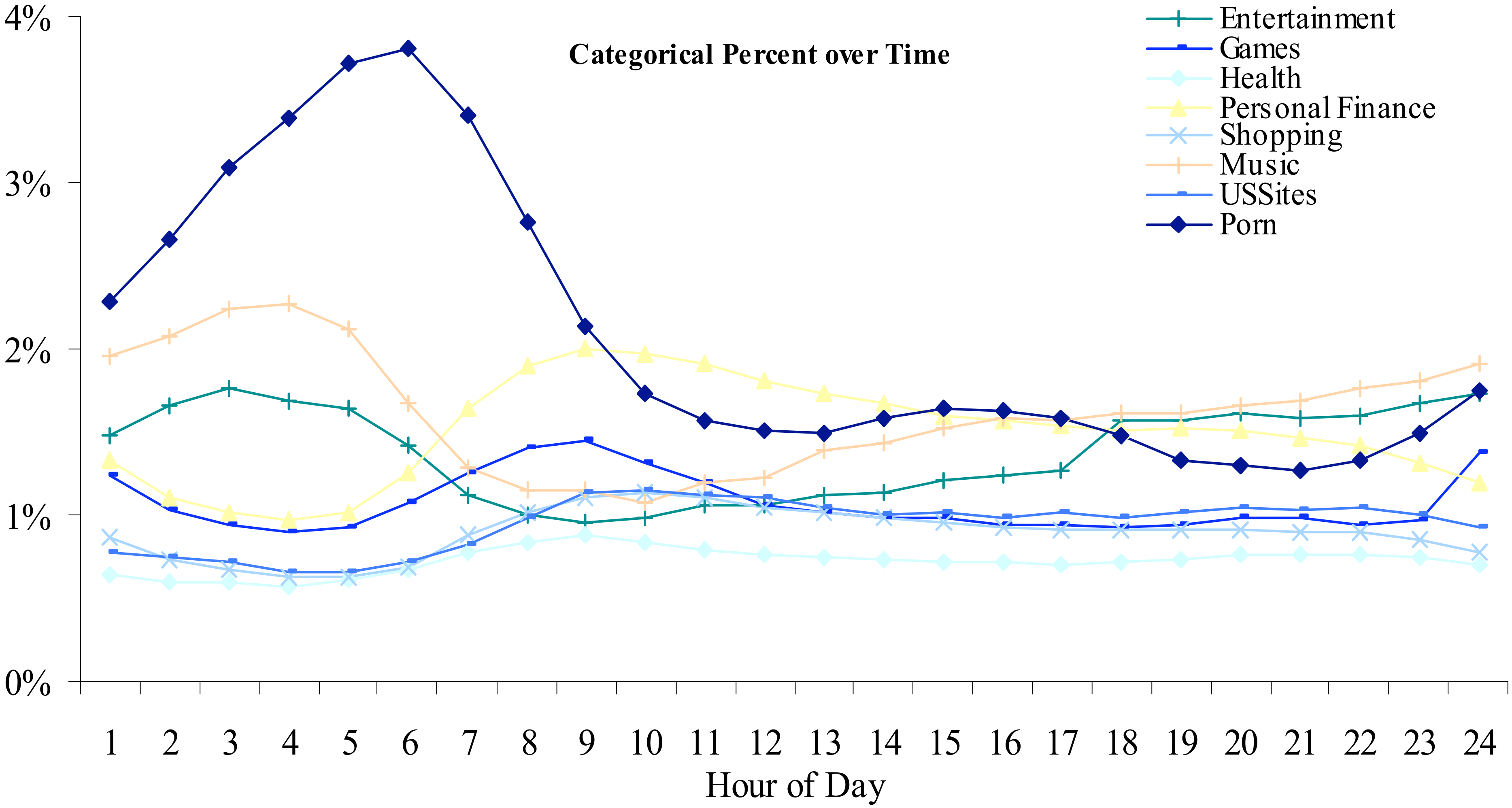
Hourly Frequency Distribution



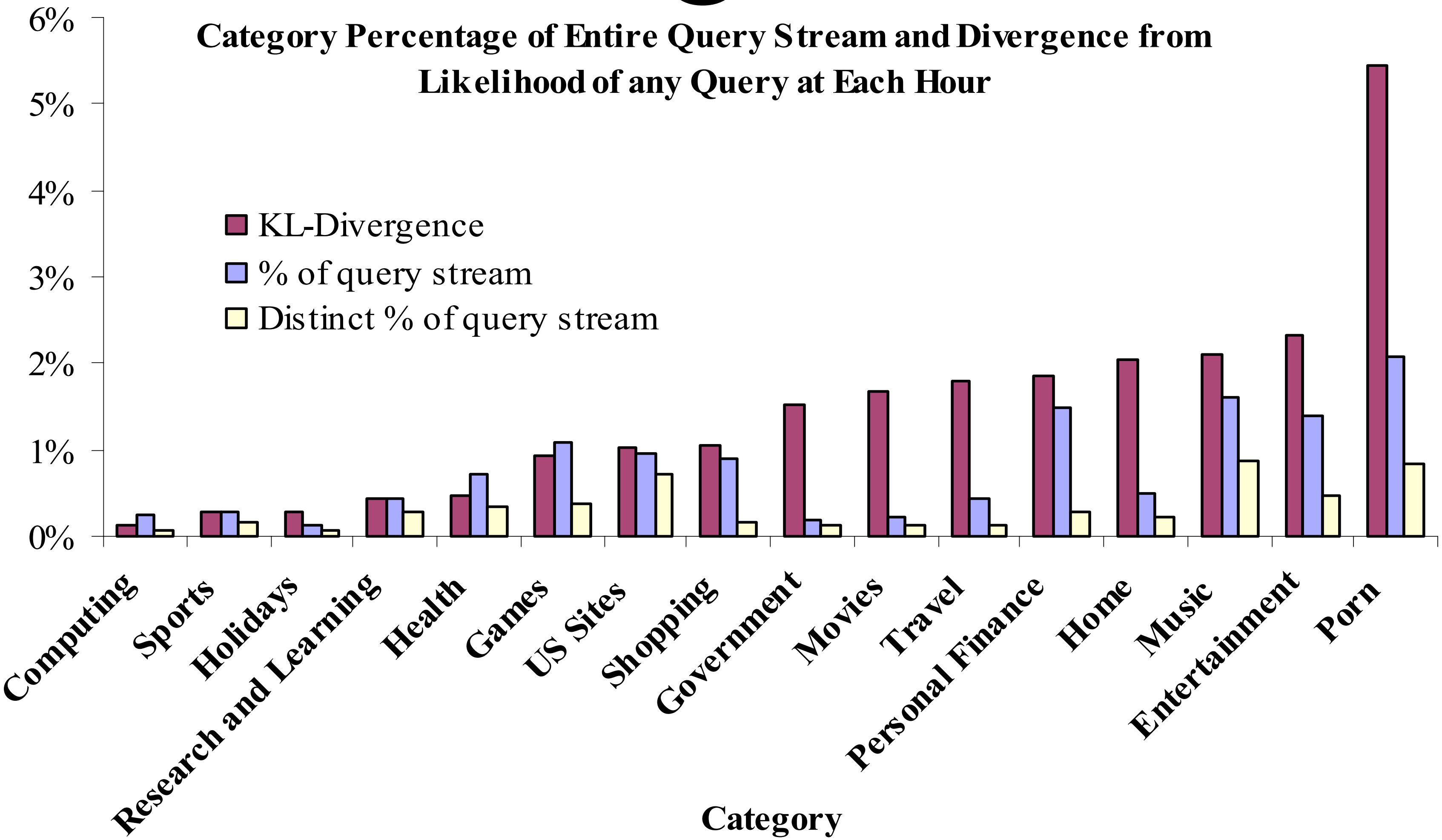
Query Categories



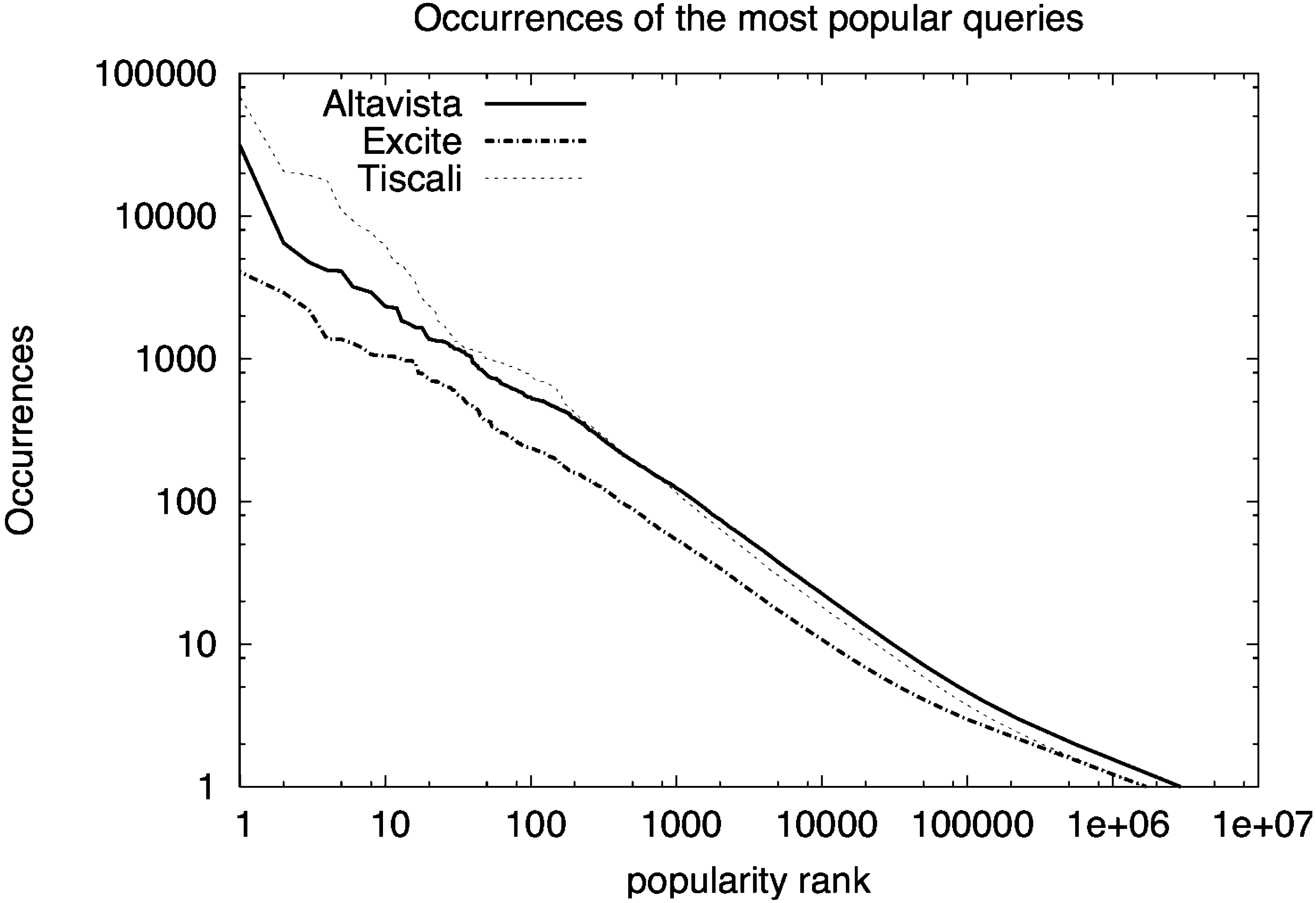
Hourly Categories



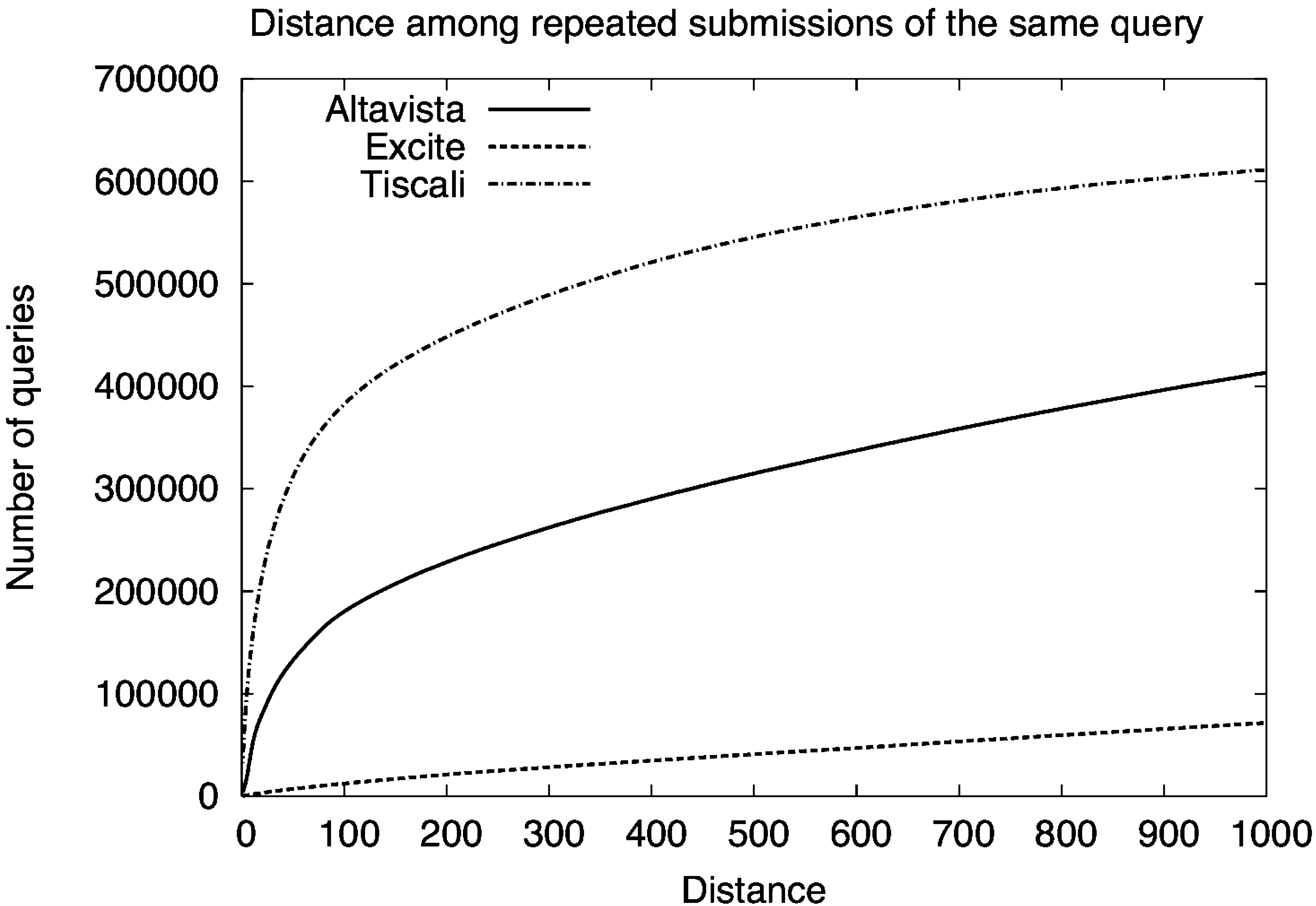
The Real Popular Categories



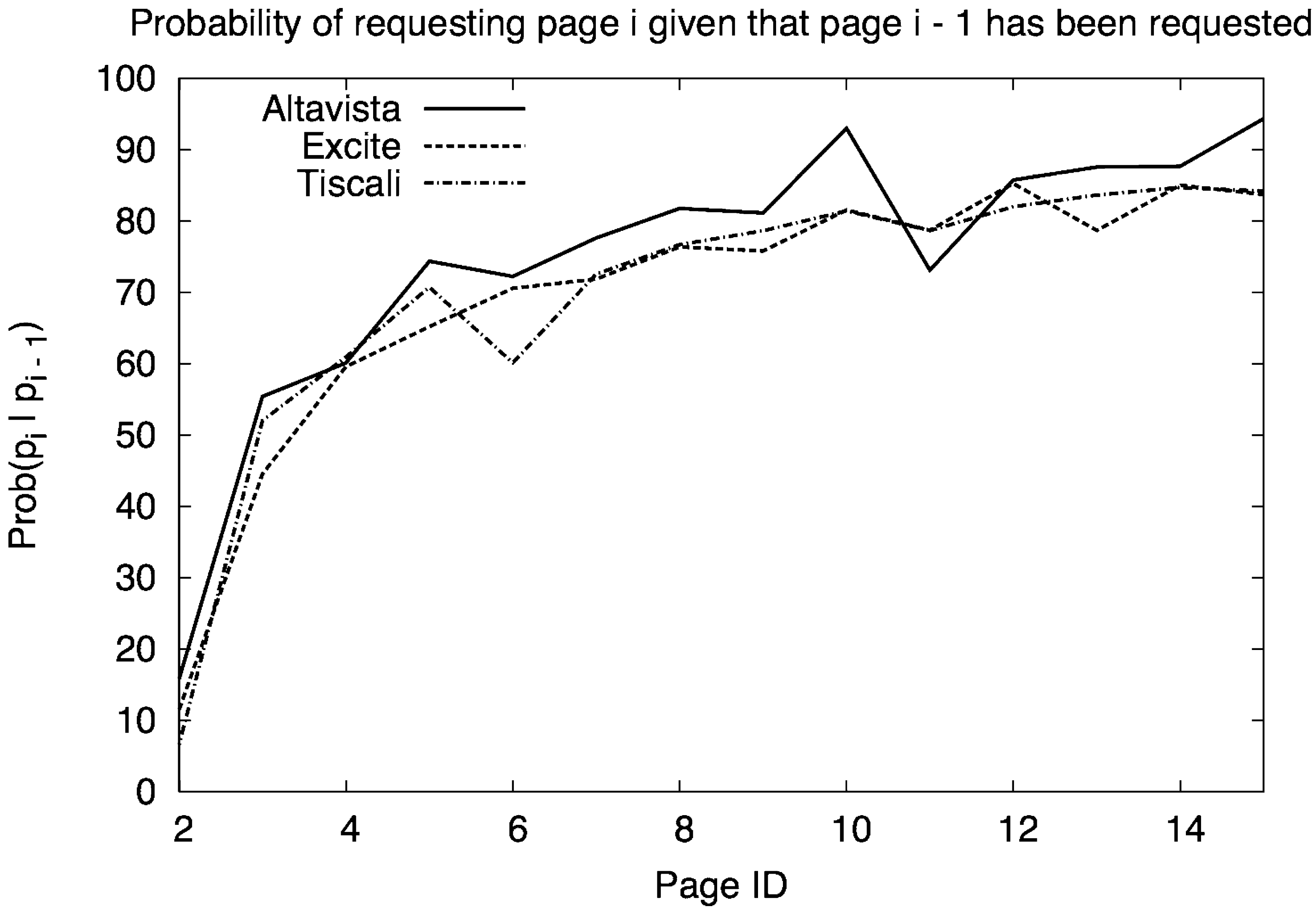
What's the OVERALL Distribution of Queries?



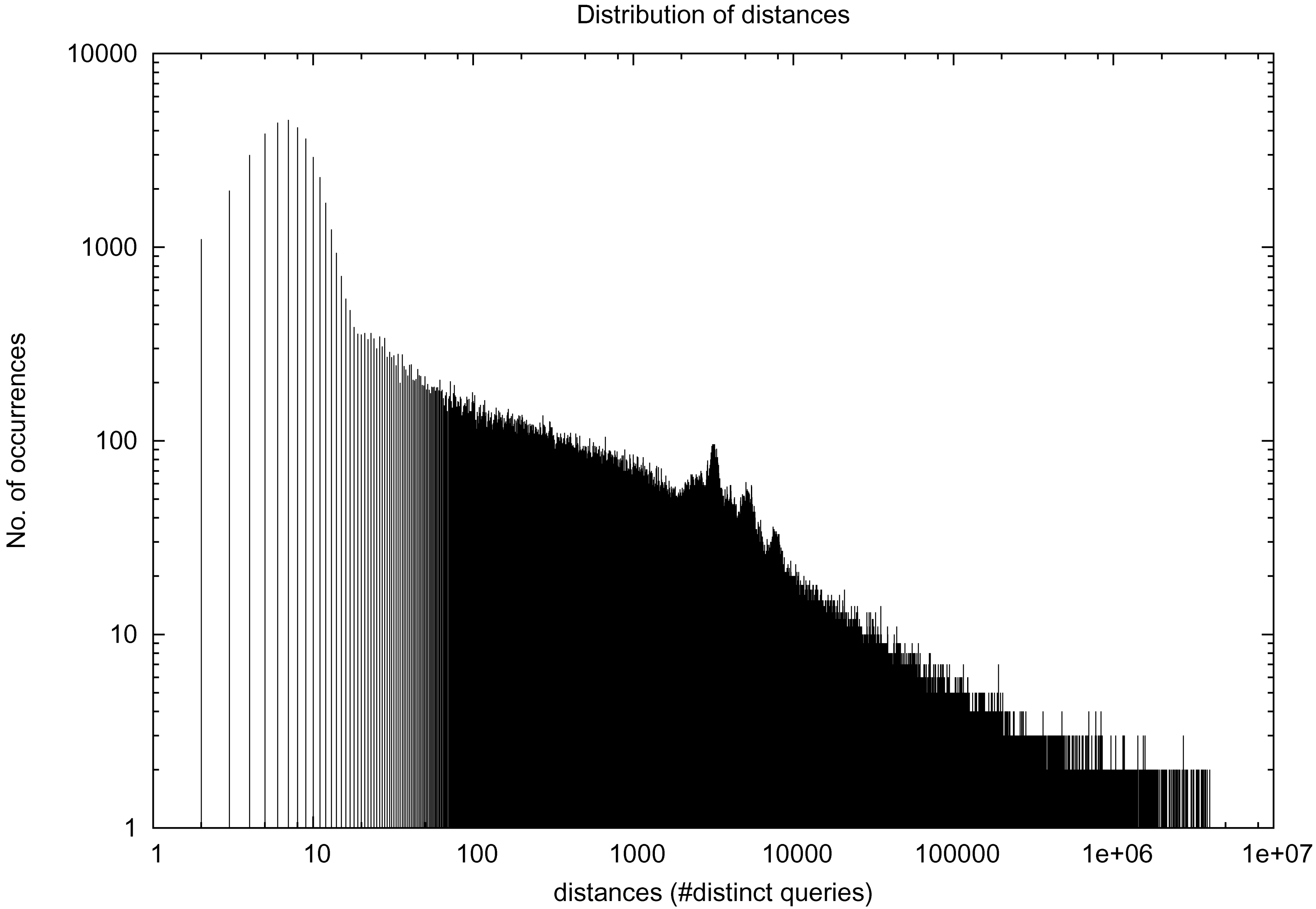
Distance Among Repetitions



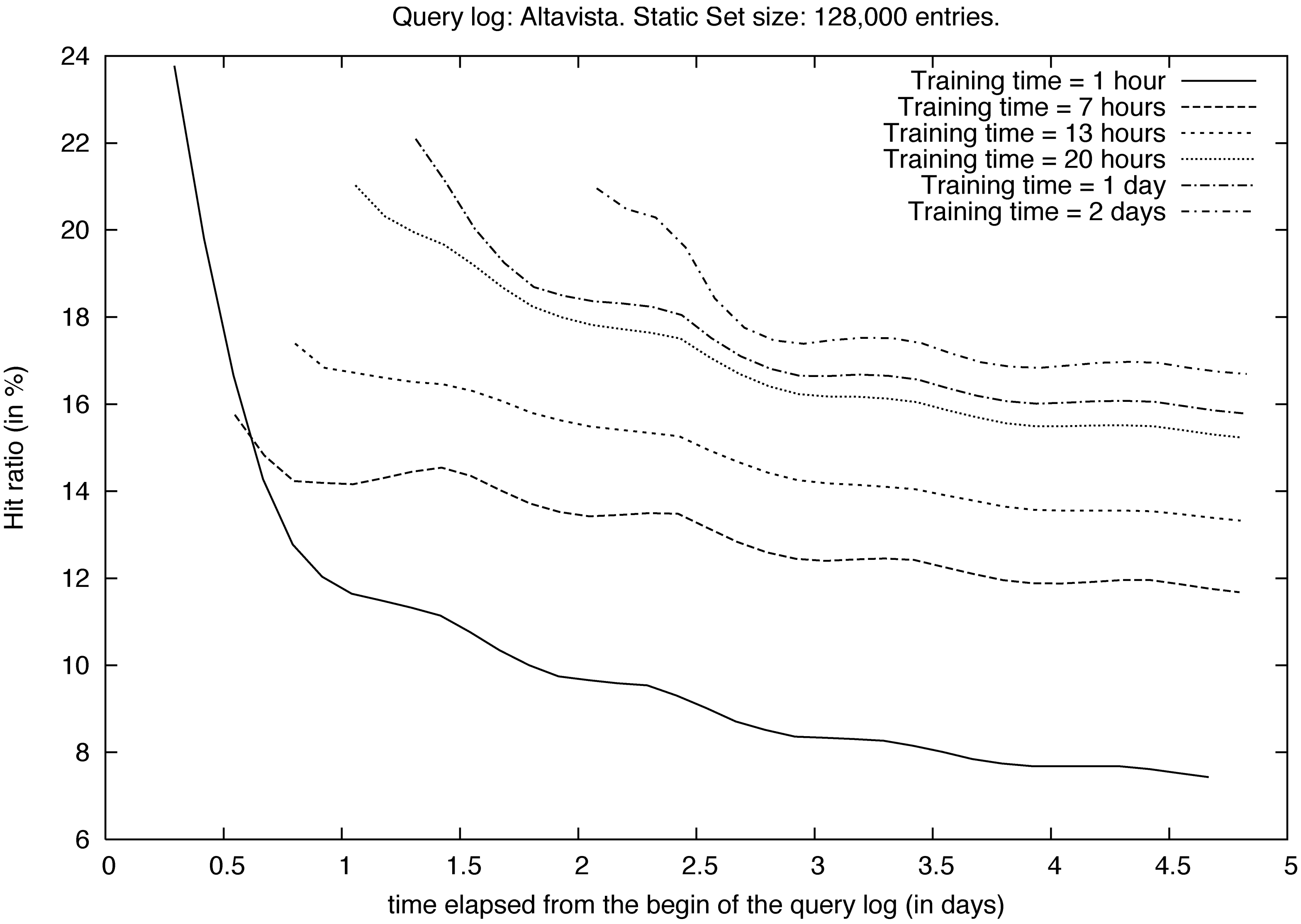
Page Request Breakdown



Distance Among Repetitions



Stability and Freshness



The Lesson is Over