

AOL search queries dataset

Autor: Jozef Tomek

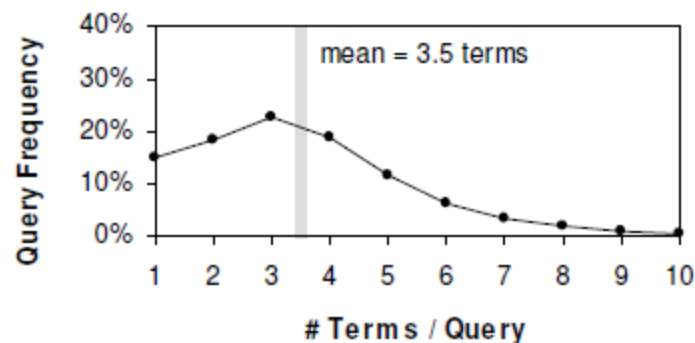
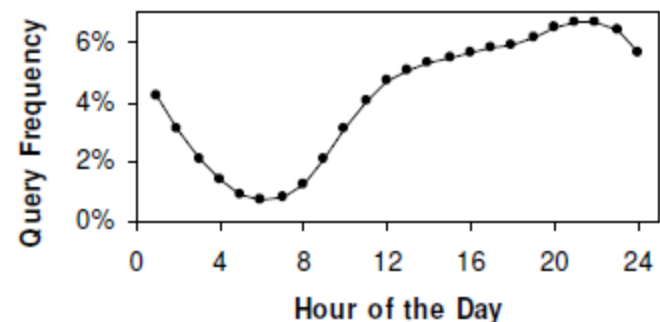
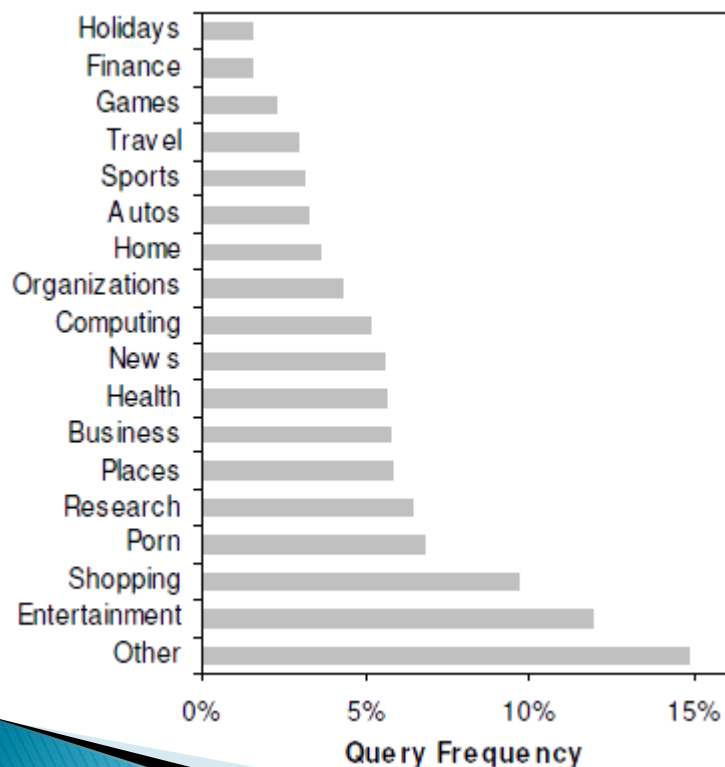
O čo sa jedná ?

- ▶ Logy reálnych dopytov na vyhľadávač AOL
- ▶ z obdobia 3. mesiacov (01.03 až 31.5 2006)
- ▶ Koľko:
 - 36,389,567 lines of data
 - 21,011,340 instances of new queries (w/ or w/o click-through)
 - 7,887,022 requests for "next page" of results
 - 19,442,629 user click-through events
 - 16,946,938 queries w/o user click-through
 - 10,154,742 unique (normalized) queries
 - 657,426 unique user ID's

Čo z toho bolo vydolované ? 1 / 4

► Query Space

- aké dopyty, ako sa menia v čase



Jozef Tomek
AOL search queries dataset

Čo z toho bolo vydolované ? 2/4

► User Sessions

- vzorce správania používateľov

(3.1)

In this session, the user formulates - and reformulates - a series of queries in pursuit of a single overall task.

28% of all queries are reformulations of a previous query. In such cases, the average query is reformulated 2.6 times.

Timeline
(mm:ss)

Query

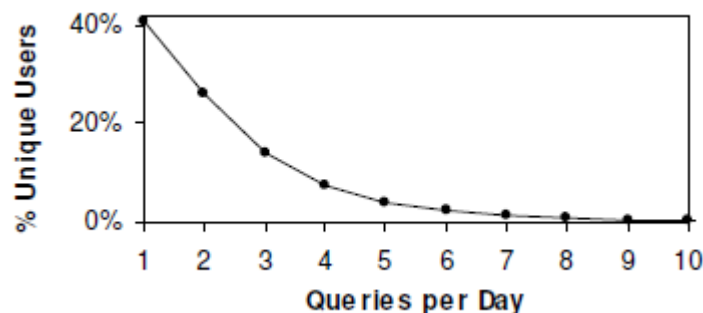
00:00	○	nursing registry
04:18	Ⓢ	certified nursing assistant 1
08:48	Ⓢ	nursing assistant registry
09:48	Ⓢ	license look up for nursing assistants
10:06	Ⓢ	nursing assistant 1 certification
11:42	Ⓢ	nursing assistant 1 license look ups
12:18	Ⓢ	nursing assistant 1 expiration look up
12:30	Ⓢ	nursing registry in Raleigh
13:24	Ⓢ	nursing aide registry of Raleigh
15:00	+	nursing aide registry of Raleigh website
16:06	<	nursing aide registry of Raleigh
19:48	Ⓢ	north carolina board of nursing information for nursing assistant 1
22:24	Ⓢ	license look up for nursing assistant 1
24:36	Ⓢ	license information for nursing assistant 1 expiration
28:30	Ⓢ	north carolina nursing assistant 1 license information

28% of all queries are reformulations of a previous query. In such cases, the average query is reformulated 2.6 times.

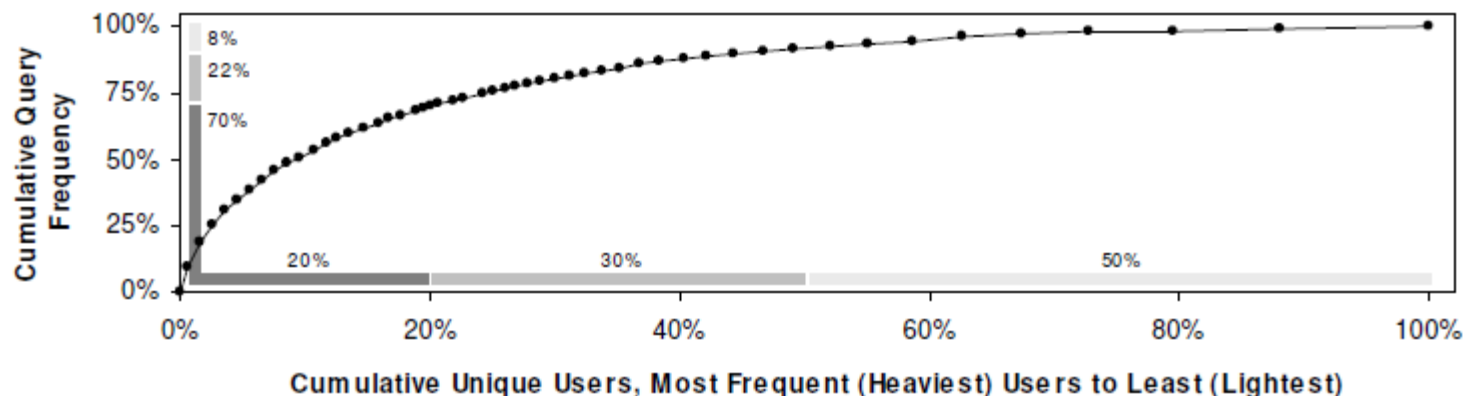
- New query
- + Add word(s) to query
- Remove word(s) from query
- Ⓢ Change word(s) in query
- > More results for same query
- < Return to a previous query
- End of session

Čo z toho bolo vydolované ? 3/4

- ▶ User behaviour
 - kam klikajú, ako veľmi vyhl'adávajú



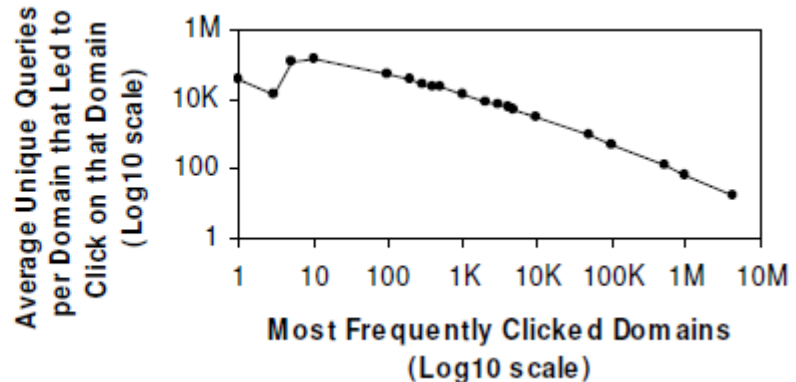
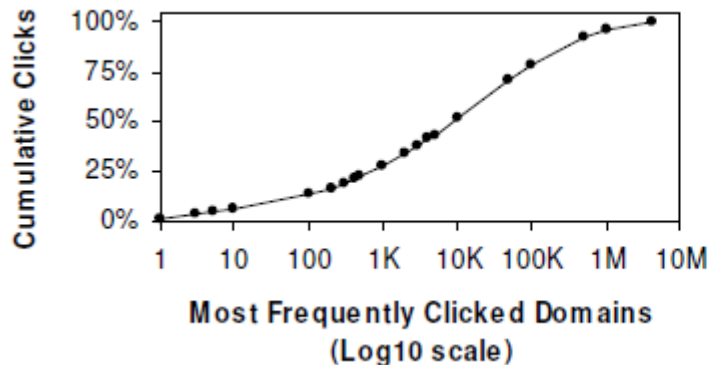
Page Rank	% Total Views	Result Rank	% Total Clicks
1	86	1	45
2	6	2	13
3	2	3	9
4	1	4	6
5	1	5	5
6	1	6	4
7	1	7	3
8	< 1	8	3
9	< 1	9	2
10	< 1	10	3
> 10	< 1	> 10	9



Jozef Tomek
AOL search queries dataset

Čo z toho bolo vydolované ? 4/4

- ▶ Content space
 - Následne navštívené domény



Forma 1 / 2

- ▶ 10 plaintext (.txt) súborov
 - ~220 MB gzipped na ~45 MB
 - spolu 440 MB
- ▶ Zorad'ované
 - **Iba v rámci každého súboru !!!**
 - podľa uniqueID používateľa
 - až potom sekvenčne (ako chodili jeho dopyty)
- ▶ Pre vlastnú potrebu som celé zoradil sekvenčne, som ochotný sa podeliť 😊

Forma 2/2

AnonID	Query	QueryTime	ItemRank	ClickURL		
6592004	farris tax services	2006-04-04 12:29:17				
6592004	carbon fiber walther	2006-04-11 11:49:35	5			http://www.maxdealer.com
6592004	carbon fiber walther	2006-04-11 12:37:08				
6592004	yelloe pages	2006-04-12 11:31:31				
6592004	interstate 40 exit 150	2006-04-12 13:47:28				
6592004	interstate 40 tennessee	2006-04-12 13:47:56				
6592004	interstate 40 tennessee	2006-04-12 13:48:44				
6592004	interstate 40 tennessee exit 154	2006-04-12 13:49:19				
6592004	bucksnort tennessee	2006-04-12 13:50:09	1			http://en.wikipedia.org
6592004	hickman county tennessee real estate	2006-04-12 13:51:53	8			http://www.relohomesearch.com
6592004	hickman county tennessee land for sale	2006-04-12 14:07:08	2			http://lyles-tn-realestate.com
6592004	7.11 acres for sale in hickman county	2006-04-12 14:12:43				
6592004	hickman county tennessee land for sale	2006-04-12 14:13:09				
6592004	interior design course	2006-04-12 16:20:53				
6592004	interior design course	2006-04-12 16:22:23				
6592004	interior design course	2006-04-12 16:24:05	27			http://www.trade-schools.net
6592004	interior design course	2006-04-12 16:24:05	28			http://www.adigitaldreamer.com
6592004	yelloe pages	2006-04-13 10:25:52				
6592004	the dixie line credit union	2006-04-13 10:27:53				

Ako sa k tomu dostať ?

► Kde?

- <http://www.gregsadetsky.com/aol-data/>
- Nie len uvedené štatistiky v
 - G. Pass, A. Chowdhury, C. Torgeson, "*A Picture of Search*" The First International Conference on Scalable Information Systems, Hong Kong, June, 2006.
 - <http://www.ir.iit.edu/~abdur/publications/pos-infoscale.pdf>
- U mňa (redemer.sk@gmail.com)

3 months
36,389,567 lines of data
21,011,340 instances of new queries (w/ or w/o click-through)
7,887,022 requests for "next page" of results
19,442,629 user click-through events
16,946,938 queries w/o user click-through
10,154,742 unique (normalized) queries
657,426 unique user ID's

Jozef Tomek
AOL search queries dataset