67-300 SEARCH ENGINES

SEARCH LOG ANALYSIS

BASED ON JAMIE CALLAN'S LECTURES

LECTURER: JOAO PALOTTI (JPALOTTI@ANDREW.CMU.EDU)

19TH APRIL 2017

LECTURE'S GOAL

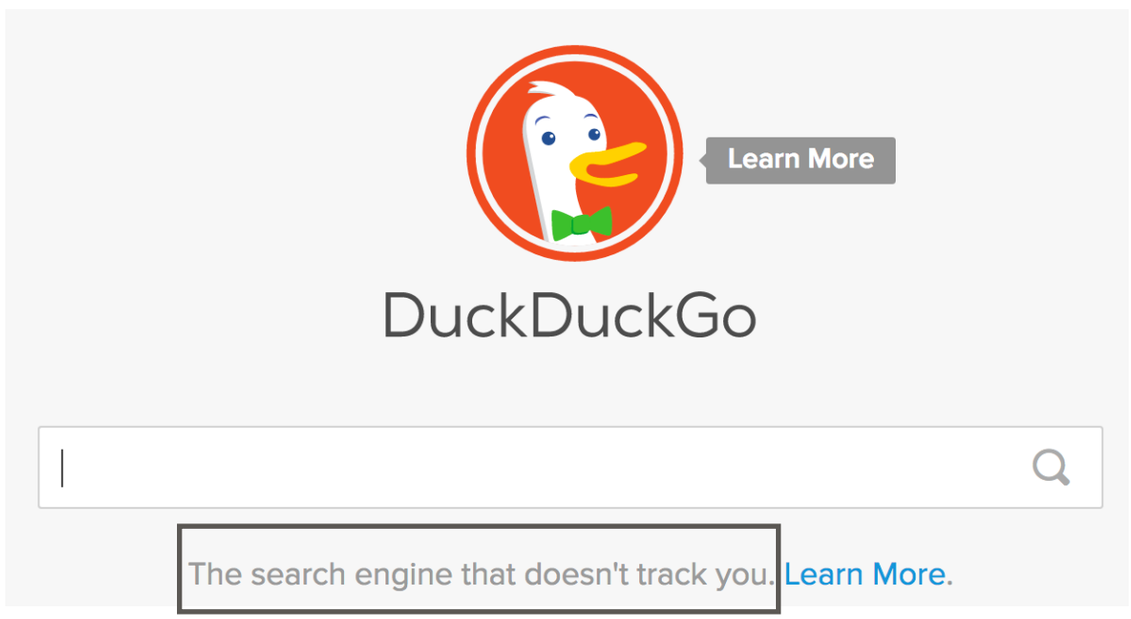
- Introduction to Search Logs
- Search Tasks:
  - Understanding user behavior
  - Segmenting search logs into sessions
  - Query Suggestions
  - Query Intents

SEARCH LOGS

- Most Web sites save information about whatever you do:
  - Search Engines are not different

SEARCH LOGS

- Most Web sites save information about whatever you do:
  - Search Engines are not different
- Few exceptions… ?



Probably…

SEARCH LOGS

- Other search engines save:
  - your query
  - a timestamp
  - your IP address
  - your session ID (in a cookie)
  - information about your operating system, browser
  - clickthrough information (which search result you clicked)

ARE SEARCH LOGS PUBLICLY AVAILABLE?

ARE SEARCH LOGS PUBLICLY AVAILABLE?

- No.

ARE SEARCH LOGS PUBLICLY AVAILABLE?

- No. **WHY?**

ARE SEARCH LOGS PUBLICLY AVAILABLE?

- No. **WHY?**
- Not available because:
  - Competitive reasons
  - Privacy reasons

ARE SEARCH LOGS PUBLICLY AVAILABLE?

- No. **WHY?**
- Not available because:
  - Competitive reasons
  - Privacy reasons
- Few logs ever publicly available. Examples:
  - The Excite log (1997): 18k users; 51k queries
  - AOL log (2006): > 650k users, >20 M queries

ARE SEARCH LOGS PUBLICLY AVAILABLE?

- No. **WHY?**
- Not available because:
  - Competitive reasons
  - Privacy reasons
- Few logs ever publicly available. Examples:
  - The Excite log (1997): 18k users; 51k queries **PRE-HISTORIC ERA OF INTERNET**
  - AOL log (2006): > 650k users, >20 M queries **PLENTY OF ISSUES**

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

ANONYMIZED SEARCH

HOW DOES AOL SEARCH LOG LOOK LIKE?

LARGE NUMBER (50%) OF QUERIES HAVE NO CLICK

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

HOW DOES AOL SEARCH LOG LOOK LIKE?

MIGHT MEAN THAT THIS IS A GOOD SEARCH ENGINE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

THEY KNOW YOUR HABITS...

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

FINANCIAL PROBLEMS?

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

SOMEONE IN CALIFORNIA WITH FINANCIAL PROBLEMS?

HOW DOES AOL SEARCH LOG LOOK LIKE?

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	staple.com	2006-03-01 21:19:29	1	staple.com
142	dfdf	2006-03-24 01:31:04		
142	-	2006-04-08 08:38:23		
217	lottery	2006-03-01 11:58:51	1	calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	ameriprise.com
...	...	...	...	...

THIS IS LITERALLY AMONG THE FIRST 50 ENTRIES OF THE WHOLE SEARCH LOG

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc... (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

PORN...

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

PORN...

BECOMING A PRIEST?

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

PORN...

BECOMING A PRIEST?

MORE PORN...

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

PORN...

BECOMING A PRIEST?

MORE PORN...

MORE PORN...

```
(ENV)[11:15:54 kronos:/data/palotti/logAnalysisDataSets/aolData/AOL-user-ct-collection]$ zgrep -F 6120607 user-ct-test-collection-01.txt.gz
6120607 bev net 2006-03-03 19:41:04     3       http://www.bevnet.com
6120607 roy rogers restaraunts  2006-03-03 19:47:18
6120607 roy rogers restaraunts franchise        2006-03-03 19:47:41
6120607 roy rogers franchise    2006-03-03 20:00:16
6120607 hot dog franchises      2006-03-03 20:02:25     2       http://www.franchisegator.com
6120607 deep fried hot dog franchise    2006-03-03 20:17:38     9       http://www.everything2.com
6120607 indian head maryland franchise opportunity      2006-03-03 20:25:08
6120607 business for sale indian head maryland  2006-03-03 20:30:11
6120607 merle allen sutphin     2006-03-03 20:33:04
6120607 whitegirlserves.com     2006-03-04 09:39:27     1       http://www.white-girl-serves.com
6120607 whitegirlserves.com     2006-03-04 09:44:38
6120607 creampiecouples 2006-03-04 09:45:41
6120607 creampieinteracial      2006-03-04 09:46:06
6120607 creampie interracial    2006-03-04 09:46:17     6       http://p097.ezboard.com
6120607 www.giantsblackmeatwhitetreat   2006-03-04 11:46:53     1       http://www.giantsblackmeatwhitetreati.com
6120607 church pulpits  2006-03-09 19:37:43
6120607 church pulpits  2006-03-09 19:38:02
6120607 church pulpits  2006-03-09 19:38:03     2       http://www.vachurchfurniture.com
6120607 church pulpits  2006-03-09 19:38:03     6       http://www.acrylicpulpits.com
6120607 church pulpits  2006-03-09 19:38:03     1       http://www.vachurchfurniture.com
6120607 pulpits 2006-03-09 19:51:07     1       http://www.christianitytoday.com
6120607 pulpits 2006-03-09 19:51:07     4       http://www.displays2go.com
6120607 pulpits 2006-03-09 19:51:07     5       http://www.catholicsupply.com
6120607 pulpits 2006-03-09 19:51:07     6       http://www.abcoffice.com
6120607 how to become a ordained pastor 2006-03-10 21:18:35
6120607 become ordained 2006-03-10 21:26:57
6120607 anal creampies  2006-03-11 01:13:40
6120607 creampie interacial     2006-03-11 01:14:49     2       http://www.forqan.com
6120607 tax secrets     2006-03-11 07:30:11
6120607 pbgcareers.com  2006-03-17 19:52:21
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:52:58     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 pbgcareers.com  2006-03-17 19:58:00     1       http://www.pbgcareers.com
6120607 bangbros        2006-03-17 23:03:17     2       http://www.bangbrosonline.com
6120607 bangbros        2006-03-17 23:03:17     10      http://galleries.bigmouthfuls.com
6120607 whitegirlserves 2006-03-17 23:12:40     2       http://www.abeservice.org
6120607 whitegirlserves 2006-03-17 23:12:40     1       http://www.white-girl-serves.com
6120607 whitegirlserves 2006-03-17 23:12:40     6       http://www.jupiterwebevens.com
6120607 blackdickswhitechicks   2006-03-17 23:17:55
6120607 youth group bible studies       2006-03-21 17:38:19     5       http://www.ministryblue.com
6120607 youth group bible studies       2006-03-21 17:38:19     1       http://www.egadideas.com
6120607 youth group bible lessons       2006-03-21 18:20:41     1       http://www.teenlifeministries.com
6120607 youth group bible lessons       2006-03-21 18:20:41     2       http://christianteens.about.com
6120607 youth group bible lessons       2006-03-21 18:20:41     3       http://www.gospelcom.net
6120607 youth group bible questions     2006-03-21 18:35:35
6120607 church youth topics     2006-03-21 18:38:24
6120607 church youth topics and lessons 2006-03-21 18:39:05     2       http://www.teenlifeministries.com
6120607 church youth topics and lessons 2006-03-21 18:39:05     8       http://www.zondervan.com
6120607 youth bible lessons     2006-03-21 18:42:28     2       http://www.teenlifeministries.com
6120607 youth bible lessons and questions       2006-03-21 18:43:55     8       http://www.newwineskin.com
```

RESTAURANTS...OKAY

PORN...

BECOMING A PRIEST?

MORE PORN...

MORE PORN...

YOUTH BIBLE LESSONS...

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc... (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc... (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:

3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 8 drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc... (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:

3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 8 drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day

2281868 how destroy demons that live in apt above
2281868 how destroy demons that live in apt above
2281868 is buddism satanism
2281868 is buddism satanism
2281868 is hip hop and rap music a form of satanism
2281868 is hip hop and rap music a form of satanism
2281868 are niggers satan or demons or gremlins
2281868 are niggers satan or demons or gremlins
2281868 livingstone college
2281868 animal sex
2281868 animal sex
2281868 animal sex
2281868 killing voyeur neighbors who are satanic cult mem
2281868 killing voyeur neighbors who are satanic cult mem
2281868 do niggers have x-ray vision
2281868 do niggers have x-ray vision

AOL SEARCH SCANDAL

2917636 knows the importance of being a good host.

2917636 date rape
2917636 is it normal to cook you rfriend breakfast after he rapes you
2917636 the morning after being raped
2917636 sexual assualt

3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 drinking 5 drinks a day
3348270 8 drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 8 alcohol drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day
3348270 drinking 10 - 15 drinks a day

2281868 how destroy demons that live in apt above
2281868 how destroy demons that live in apt above
2281868 is buddism satanism
2281868 is buddism satanism
2281868 is hip hop and rap music a form of satanism
2281868 is hip hop and rap music a form of satanism
2281868 are niggers satan or demons or gremlins
2281868 are niggers satan or demons or gremlins
2281868 livingstone college
2281868 animal sex
2281868 animal sex
2281868 animal sex
2281868 killing voyeur neighbors who are satanic cult mem
2281868 killing voyeur neighbors who are satanic cult mem
2281868 do niggers have x-ray vision
2281868 do niggers have x-ray vision

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc... (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:
- NYT report could identify at least a user, Thelma Arnold (#4417749), a 62-year-old widow from Lilburn, Georgia

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc… (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:
- NYT report could identify at least a user, Thelma Arnold (#4417749), a 62-year-old widow from Lilburn, Georgia
- Some people got fired from AOL… (as the CTO of twitter)

AOL SEARCH SCANDAL

- Internet groups created to discover who are those people, what they are looking for, etc… (not researchers)
- Many “fun”/“dangerous”/“disturbing” links on the results:
- NYT report could identify at least a user, Thelma Arnold (#4417749), a 62-year-old widow from Lilburn, Georgia
- Some people got fired from AOL… (as the CTO of twitter)

FOMENT RESEARCH IN MANY WAYS….

HIGHLY RECOMMENDED PAPER

http://dl.acm.org/citation.cfm?id=1146848

A Picture of Search

Greg Pass
America Online
gregpass1@aol.com

Abdur Chowdhury
America Online
cabdur@aol.com

Cayley Torgeson
Raybeam
torgeson@raybeam.com

A Picture of Search

Greg Pass
America Online
gregpass1@aol.com

Abdur Chowdhury
America Online
cabdur@aol.com

Cayley Torgeson
Raybeam
torgeson@raybeam.com

ABSTRACT

We survey many of the measures used to describe and evaluate the efficiency and effectiveness of large-scale search services. These measures, herein visualized versus verbalized, reveal a domain rich in complexity and scale. We cover six principle facets of search: the query space, users' query sessions, user behavior, operational requirements, the content space, and user demographics. While this paper focuses on measures, the measurements themselves raise questions and suggest avenues of further investigation.

Keywords: system modeling, user modeling, distributed database searching, search methods, user interfaces.

1. INTRODUCTION

Large-scale search services, such as Yahoo and Google, index billions of pages of content in order to service billions of user queries. In order to maintain tractability in this highly scaled environment, operators of such services use a number of measures to evaluate the ongoing efficiency (e.g., user latency) and effectiveness (e.g., search result precision) of their systems. We survey a number of these measures – in particular, measures that we, as operators ourselves of a large-scale search service, have found to be descriptive and useful.

We organize these measures into six principle facets of a large-scale search service, and the following six sections explore each facet in turn. They are: Section 2, Query Space, which describes the population of user queries, and, in particular, how those queries change over time; Section 3, User Sessions, which describes the pattern of query formulations users express within the scope of single sessions; Section 4, User Behavior, which describes populations of users' interactions with the search service, with clickthrough, as one trace of user interaction, given particular focus; Section 5, Operational Requirements, which describes the runtime efficiency of a search service; Section 6, Content Space, which describes the population of search results, and the content those results represent, serviced by search services; and Section 7, User Demographics, which highlights the geographics of demographics.

In each section, the graphical measures themselves comprise the majority of the sectional content[1], with supplementary text given in the form of either graphic annotations or short summaries of the section as a whole. We have chosen this style of presentation for several reasons. Foremost, effective measures, as essential vehicles of large-scale tractability, should speak for themselves: it is in their best interests, for, ultimately, these measures, sometimes directly, sometimes indirectly, are the operators' only quantitative handle on the quality of the search service. Presenting the measures graphically – and densely, as a single page per section – also aids the reader in appreciating the relationships between measures, and eases holistic ruminations.

Many of the measures and measurements so presented raise additional questions. In some cases, our presentation is simply incomplete, as we have surveyed measures broadly, across six distinct facets of search. In most cases, however, these questions will address topics requiring further investigation, and we hope the data presented in this paper will encourage such pursuits.

[1] To assure legibility, this paper requires a 600 dpi (or greater) printer.

INFOSCALE '06. Proceedings of the First International Conference on Scalable Information Systems, May 29-June 1 2006, Hong Kong

HIGHLY RECOMMENDED PAPER

http://dl.acm.org/citation.cfm?id=1146848

A Picture of Search

Greg Pass
America Online
gregpass1@aol.com

Abdur Chowdhury
America Online
cabdur@aol.com

Cayley Torgeson
Raybeam
torgeson@raybeam.com

Not all the scientific papers have to be boring…

A Picture of Search

Greg Pass
America Online
gregpass1@aol.com

Abdur Chowdhury
America Online
cabdur@aol.com

Cayley Torgeson
Raybeam
torgeson@raybeam.com

ABSTRACT

We survey many of the measures used to describe and evaluate the efficiency and effectiveness of large-scale search services. These measures, herein visualized versus verbalized, reveal a domain rich in complexity and scale. We cover six principle facets of search: the query space, users' query sessions, user behavior, operational requirements, the content space, and user demographics. While this paper focuses on measures, the measurements themselves raise questions and suggest avenues of further investigation.

Keywords: system modeling, user modeling, distributed database searching, search methods, user interfaces.

1. INTRODUCTION

Large-scale search services, such as Yahoo and Google, index billions of pages of content in order to service billions of user queries. In order to maintain tractability in this highly scaled environment, operators of such services use a number of measures to evaluate the ongoing efficiency (e.g., user latency) and effectiveness (e.g., search result precision) of their systems. We survey a number of these measures – in particular, measures that we, as operators ourselves of a large-scale search service, have found to be descriptive and useful.

We organize these measures into six principle facets of a large-scale search service, and the following six sections explore each facet in turn. They are: Section 2, Query Space, which describes the population of user queries, and, in particular, how those queries change over time; Section 3, User Sessions, which describes the pattern of query formulations users express within the scope of single sessions; Section 4, User Behavior, which describes populations of users' interactions with the search service, with clickthrough, as one trace of user interaction, given particular focus; Section 5, Operational Requirements, which describes the runtime efficiency of a search service; Section 6, Content Space, which describes the population of search results, and the content those results represent, serviced by search services; and Section 7, User Demographics, which highlights the geographics of demographics.

In each section, the graphical measures themselves comprise the majority of the sectional content[1], with supplementary text given in the form of either graphic annotations or short summaries of the section as a whole. We have chosen this style of presentation for several reasons. Foremost, effective measures, as essential vehicles of large-scale tractability, should speak for themselves: it is in their best interests, for, ultimately, these measures, sometimes directly, sometimes indirectly, are the operators' only quantitative handle on the quality of the search service. Presenting the measures graphically – and densely, as a single page per section – also aids the reader in appreciating the relationships between measures, and eases holistic ruminations.

Many of the measures and measurements so presented raise additional questions. In some cases, our presentation is simply incomplete, as we have surveyed measures broadly, across six distinct facets of search. In most cases, however, these questions will address topics requiring further investigation, and we hope the data presented in this paper will encourage such pursuits.

[1] To assure legibility, this paper requires a 600 dpi (or greater) printer.

INFOSCALE '06. Proceedings of the First International Conference on Scalable Information Systems, May 29-June 1 2006, Hong Kong

HIGHLY RECOMMENDED PAPER

http://dl.acm.org/citation.cfm?id=1146848

Not all the scientific papers have to be boring…

A Picture of Search

Greg Pass
America Online
gregpass1@aol.com

Abdur Chowdhury
America Online
cabdur@aol.com

Cayley Torgeson
Raybeam
torgeson@raybeam.com

ABSTRACT

We survey many of the measures used to describe and evaluate the efficiency and effectiveness of large-scale search services. These measures, herein visualized versus verbalized, reveal a domain rich in complexity and scale. We cover six principle facets of search: the query space, users' query sessions, user behavior, operational requirements, the content space, and user demographics. While this paper focuses on measures, the measurements themselves raise questions and suggest avenues of further investigation.

Keywords: system modeling, user modeling, distributed database searching, search methods, user interfaces.

1. INTRODUCTION

Large-scale search services, such as Yahoo and Google, index billions of pages of content in order to service billions of user queries. In order to maintain tractability in this highly scaled environment, operators of such services use a number of measures to evaluate the ongoing efficiency (e.g., user latency) and effectiveness (e.g., search result precision) of their systems. We survey a number of these measures – in particular, measures that we, as operators ourselves of a large-scale search service, have found to be descriptive and useful.

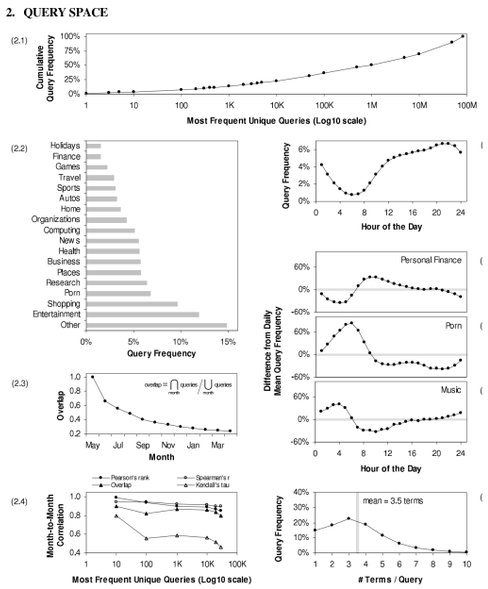
We organize these measures into six principle facets of a large-scale search service, and the following six sections explore each facet in turn. They are: Section 2, Query Space, which describes the population of user queries, and, in particular, how those queries change over time; Section 3, User Sessions, which describes the pattern of query formulations users express within the scope of single sessions; Section 4, User Behavior, which describes populations of users' interactions with the search service, with clickthrough, as one trace of user interaction, given particular focus; Section 5, Operational Requirements, which describes the runtime efficiency of a search service; Section 6, Content Space, which describes the population of search results, and the content those results represent, serviced by search services; and Section 7, User Demographics, which highlights the geographics of demographics.

In each section, the graphical measures themselves comprise the majority of the sectional content[1], with supplementary text given in the form of either graphic annotations or short summaries of the section as a whole. We have chosen this style of presentation for several reasons. Foremost, effective measures, as essential vehicles of large-scale tractability, should speak for themselves: it is in their best interests, for, ultimately, these measures, sometimes directly, sometimes indirectly, are the operators' only quantitative handle on the quality of the search service. Presenting the measures graphically – and densely, as a single page per section – also aids the reader in appreciating the relationships between measures, and eases holistic ruminations.

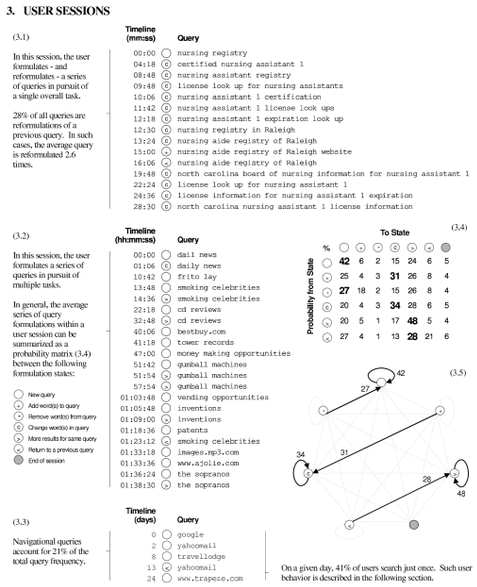
Many of the measures and measurements so presented raise additional questions. In some cases, our presentation is simply incomplete, as we have surveyed measures broadly, across six distinct facets of search. In most cases, however, these questions will address topics requiring further investigation, and we hope the data presented in this paper will encourage such pursuits.

[1] To assure legibility, this paper requires a 600 dpi (or greater) printer.

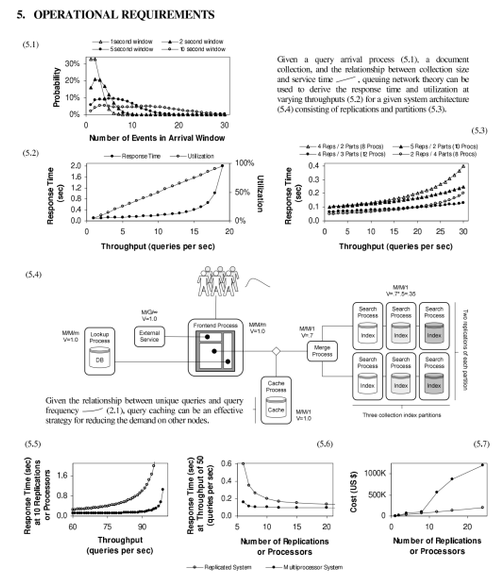
INFOSCALE '06. Proceedings of the First International Conference on Scalable Information Systems, May 29-June 1 2006, Hong Kong



The query space is vast (2.1), topically diverse (2.2), and constantly changing (2.3 - 2.8). This complexity of scale is the product of just 3.5 words per query (2.9), e



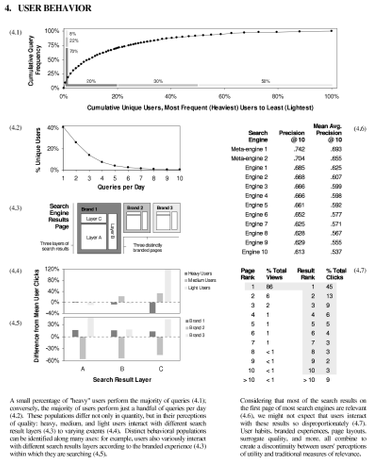
On a given day, 41% of users search just once. Such user behavior is described in the following section.



Given a query arrival process (5.1), a document collection, and the relationship between collection size and service time , queuing network theory can be used to derive the response time and utilization at varying throughputs (5.2) for a given system architecture (5.4) consisting of replications and partitions (5.3).

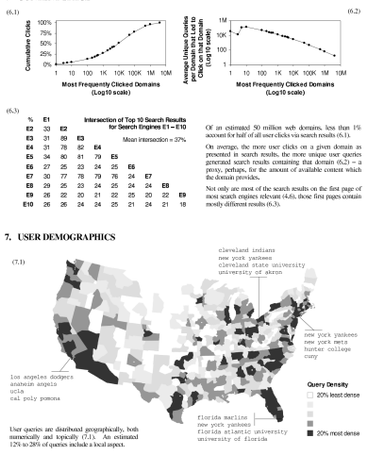
Given the relationship between unique queries and query frequency (2.1), query caching can be an effective strategy for reducing the demand on other nodes.

Multiprocessor systems can optimize queue distributions for significant performance gains (5.5, 5.6). Trends in hardware costs, though, suggest a different solution (5.7).



A small percentage of "heavy" users perform the majority of queries (4.1); conversely, the majority of users perform just a handful of queries per day (4.2). These populations differ not only in quantity, but in their perceptions of quality: heavy, medium, and light users interact with different search result layers (4.3) to varying extents (4.4). Distinct behavioral populations can be identified along many axes: for example, users also variously interact with different search results layers according to the branded experience (4.3) within which they are searching (4.5).

Considering that most of the search results on the first page of most search engines are relevant (4.6), we might not expect that users interact with these results so disproportionately (4.7). User habits, branded experiences, page layouts, surrogate quality, and more, all combine to create a discontinuity between users' perceptions of utility and traditional measures of relevance.



Of an estimated 50 million web domains, less than 1% account for half of all user clicks via search results (6.1).

On average, the more user clicks on a given domain as presented in search results, the more unique user queries generated search results containing that domain (6.2) – a proxy, perhaps, for the amount of available content which the domain provides.

Not only are most of the search results on the first page of most search engines relevant (4.6), those first pages contain mostly different results (6.3).

7. USER DEMOGRAPHICS

User queries are distributed geographically, both numerically and topically (7.1). An estimated 12% to 28% of queries include a local aspect.

8. CONCLUSION

Altogether, the measures presented convey a problem domain of significant complexity and scale; within this measured complexity, though, arise many opportunities for further research. For example, the session probability matrix (3.4) could be used to distill common patterns of query reformulations that may be helpful to users; the identification of user behavior populations in relation to page layer clickthrough (4.4) could be used to dynamically rearrange page layers with respect to that particular population; the result clickthrough distribution (4.7) as disproportionate to result relevance (4.6) may encourage continued thought on alternative measures of result utility; the query density map (7.1) could be factored into optimal data center placements; and so forth. We hope this survey will, in some measure, encourage such future investigations.

ACKNOWLEDGEMENTS

For their assistance and helpful suggestions, we thank Sudhir Achuthan, Sharon Rose, Mike Hayes, and Jay Virdy, all from America Online.

FIGURE NOTES

Figure 2.2 was presented as a pie chart in [1, 2]; we presented the same as a bar chart for clarity. Figures 2.3 and 2.4 were presented in [2, 3]; we re-specified the month axis in 2.3 to span a complete year, and updated 2.4 to use a log scale. Figures 2.6, 2.7, and 2.8 were taken from [4, 5], with the y-axis given in absolute units; we converted this axis to use a mean difference scale.

To explore Belkin's ASK hypothesis [6], we presented an example, in figure 3.1, of the difficulty some users have formulating a query in pursuit of a given task. In figures 3.4 and 3.5, we presented the user reformulation session data from [7] in the form of a state transition probability matrix. Prior work on user sessions [8] typically defines time-based session boundaries; to show the inherent ambiguity of this approach, we presented figure 3.2, in which a user pursues concurrent tasks within the same timeframe. Figure 3.3 highlights the fact that some user queries are navigational in intent, not informational; our own estimate of 21% is in line with prior work [9, 10].

In prior unpublished work, we constructed a test scenario using ten commercial web search engines and two web meta-search in order to examine large-scale repeatability of effectiveness evaluations on the web; these results were presented in Figure 4.6. This approach is tied closely to prior work in [11, 12].

Figures 5.2, 5.3, 5.5, 5.6, and 5.7 are reprints from [13]. Figure 5.1 examined the arrival process for web servers in a large-scale search service, an aspect mentioned but not presented in [13]. Figure 5.4 is also based upon [13], modified to directly relate to the figures presented in Section 4.

Although several prior works discuss the size of the web [14, 15], none present the relative distribution of web pages with respect to user clicks, as in figure 6.1, or to user queries, as in figure 6.2. The data in figure 6.3 was taken from [11] to show the same with respect to search engines.

In figure 7.1, the 12% estimate is drawn from prior work [16], while the 28% estimate is based upon an AOL internal study. There is no established definition of what it means for a query to include a "local" aspect.

REFERENCES

[1] S. Beitzel, E. Jensen, D. Lewis, A. Chowdhury, A. Kolcz, and O. Frieder, "Improving Automatic Query Classification via Semi-supervised Learning," presented at The Fifth IEEE International Conference on Data Mining, New Orleans, Louisiana, U.S.A., 2005.

[2] A. Chowdhury, "Automatic Evaluation of Web Search Services," in *Advances in Computers*, vol. ISBN: 0-12-012164-6, 2005.

[3] W. Xi, K. Sidhu, and A. Chowdhury, "Latent Query Stability," presented at International Conference on Information and Knowledge Engineering IKE, 2003.

[4] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly Analysis of a Very Large Topically Categorized Web Query Log," presented at ACM-SIGIR, 2004.

[5] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly Analysis of a Very Large Topically Categorized Web Query Log," *SIGIR*, vol. to appear, 2006.

[6] N. Belkin, R. Oddy, and H. Brooks, "ASK for information retrieval. Part I: Background and theory; Part II: Results of a design study," *Journal of Documentation*, vol. 38, pp. 61-71, 145-164, 1982.

[7] G. Murray, J. Lin, and A. Chowdhury, "Characterizing Web Search User "Sessions" with Hierarchical Agglomerative Clustering," presented at under review, 2006.

[8] D. He, A. Göker, and D. Harper, "Combining Evidence for Automatic Web Session Identification," *Information Processing & Management* vol. 38, pp. 727-742, 2002.

[9] A. Broder, "A taxonomy of web search," presented at SIGIR Forum, 2002.

[10] D. Rose and D. Levinson, "Understanding User Goals in Web Search," presented at WWW, 2004.

[11] E. Jensen, S. Beitzel, A. Chowdhury, and O. Frieder, "A Framework for Determining Necessary Query Set Sizes to Evaluate Web Search Effectiveness," presented at WWW, 2005.

[12] S. Beitzel, E. Jensen, A. Chowdhury, G. Pass, and O. Frieder, "Surrogate Scoring for Improved Measurement Precision," presented at Proceedings of the 2005 ACM Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005.

[13] A. Chowdhury and G. Pass, "Operational Requirements for Scalable Search Systems," presented at ACM-CIKM Conference for Information and Knowledge Management, 2003.

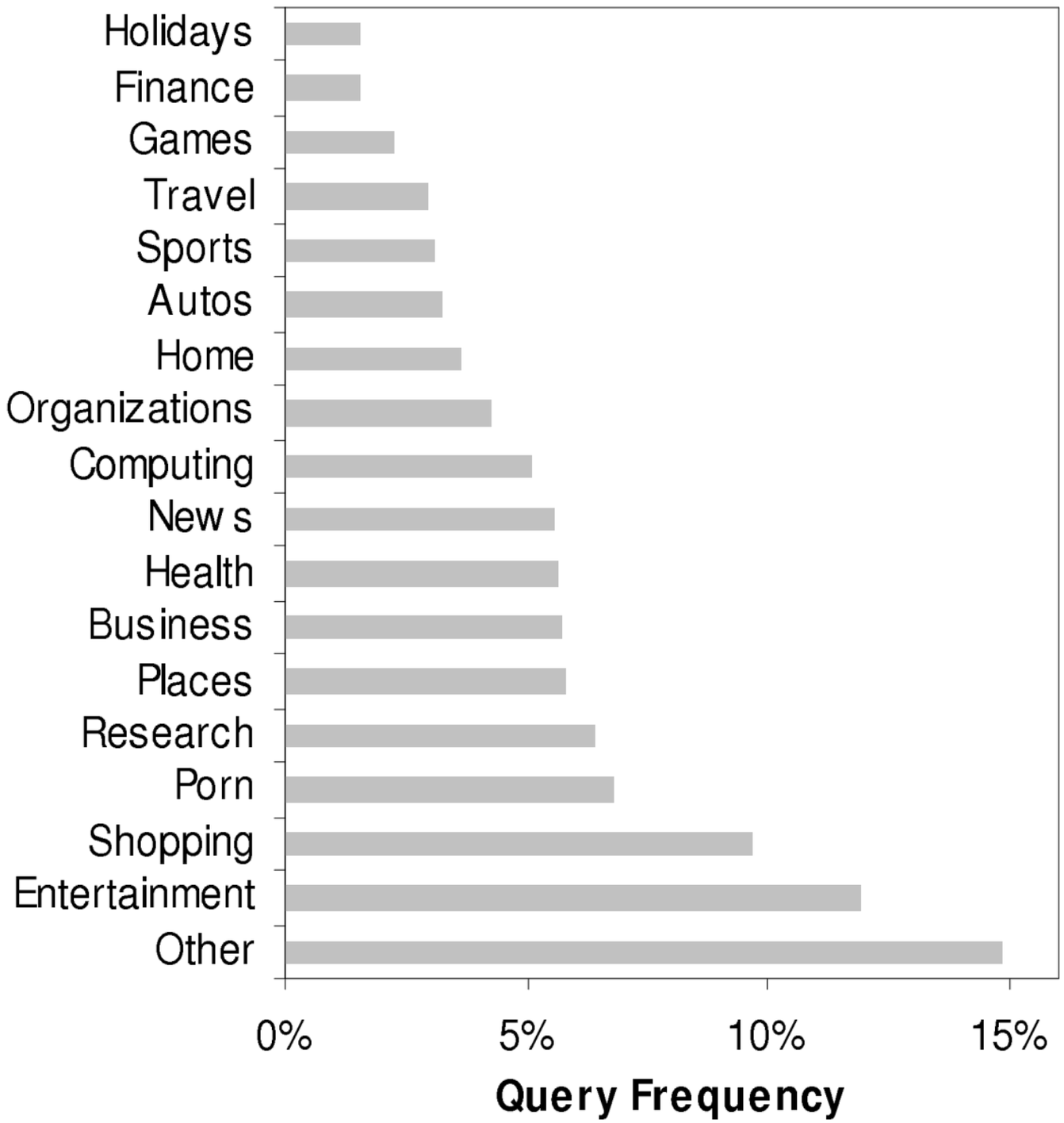
[14] A. Gulli and A. Signorini, "The Indexable Web is More than 11.5 Billion Pages," presented at WWW, 2005.

[15] P. Boutin, "Brewster Kahle made a copy of the Internet. Now, he wants your files," vol. 2005: Slate, 2005.

[16] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein, "Categorizing Web Queries According to Geographical Locality," presented at ACM-CIKM, 2003.

UNDERSTANDING SEARCHES/SEARCHERS

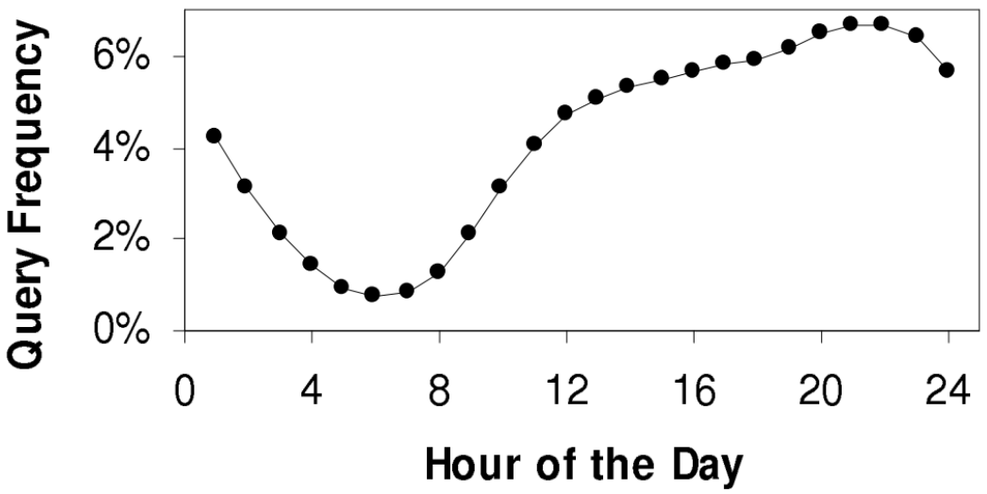
http://dl.acm.org/citation.cfm?id=1146848



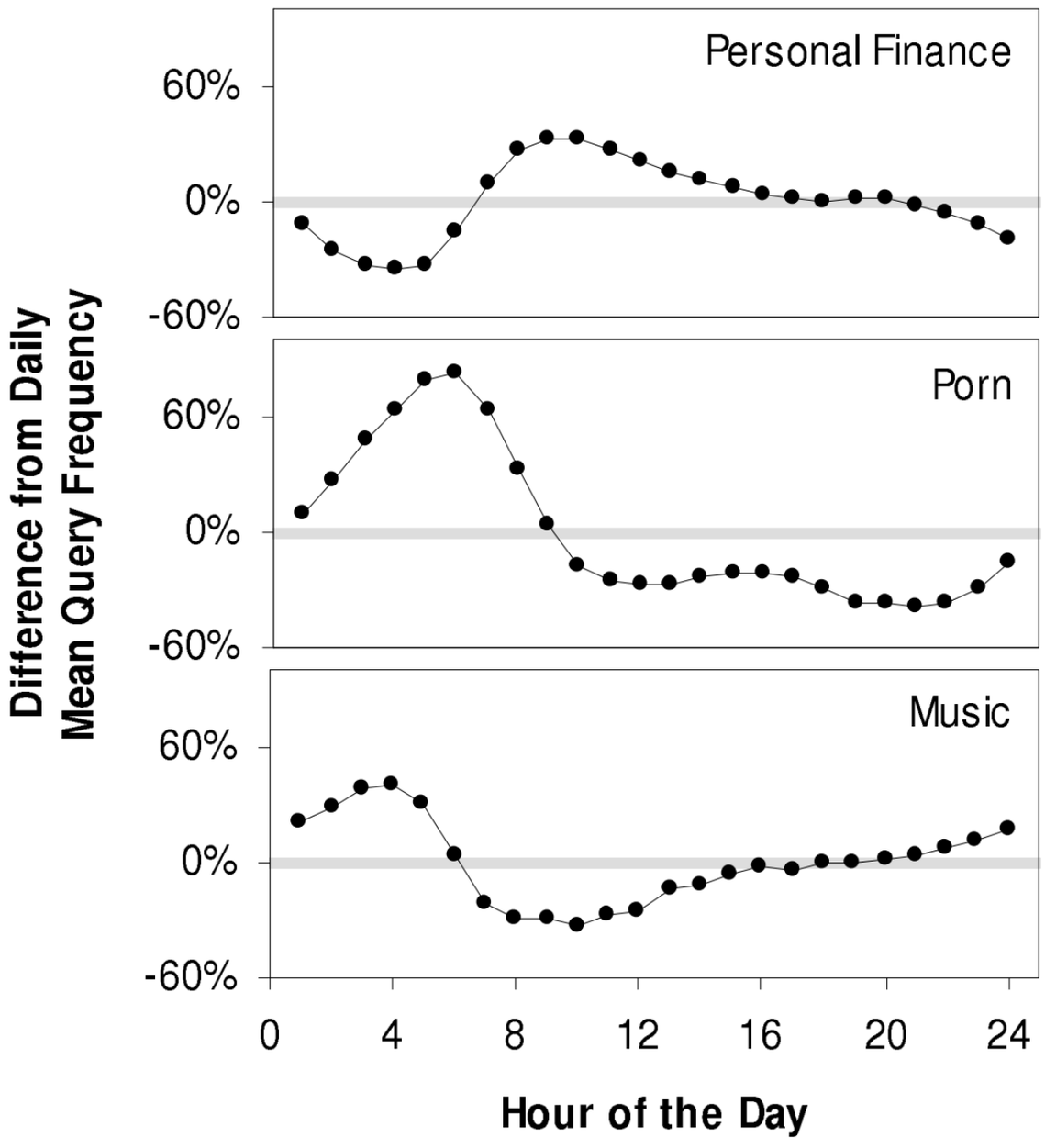
WHAT?

UNDERSTANDING SEARCHES/SEARCHERS

http://dl.acm.org/citation.cfm?id=1146848

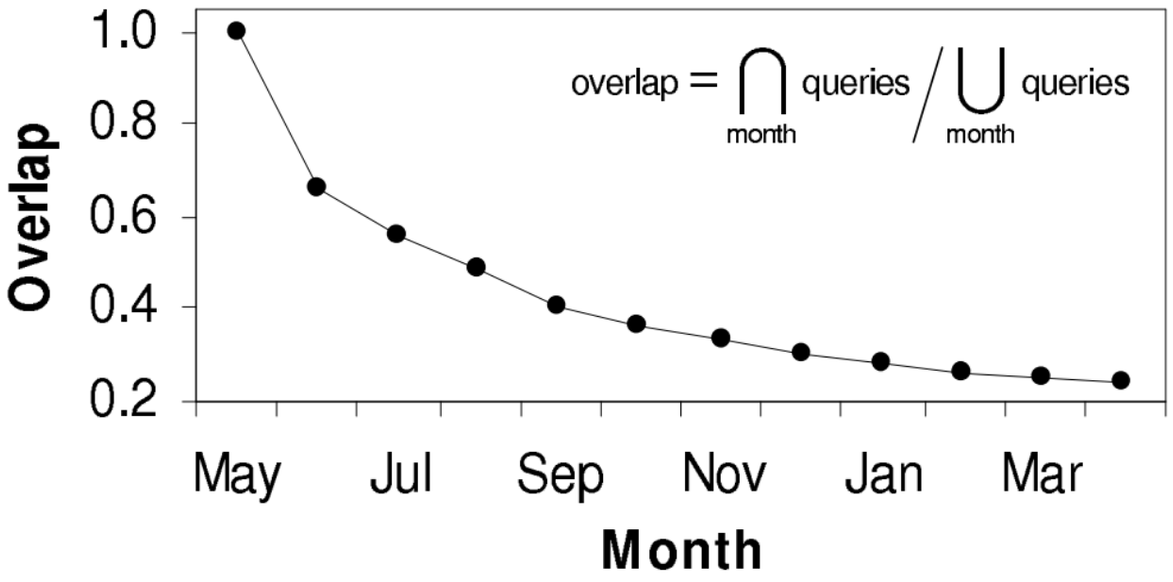


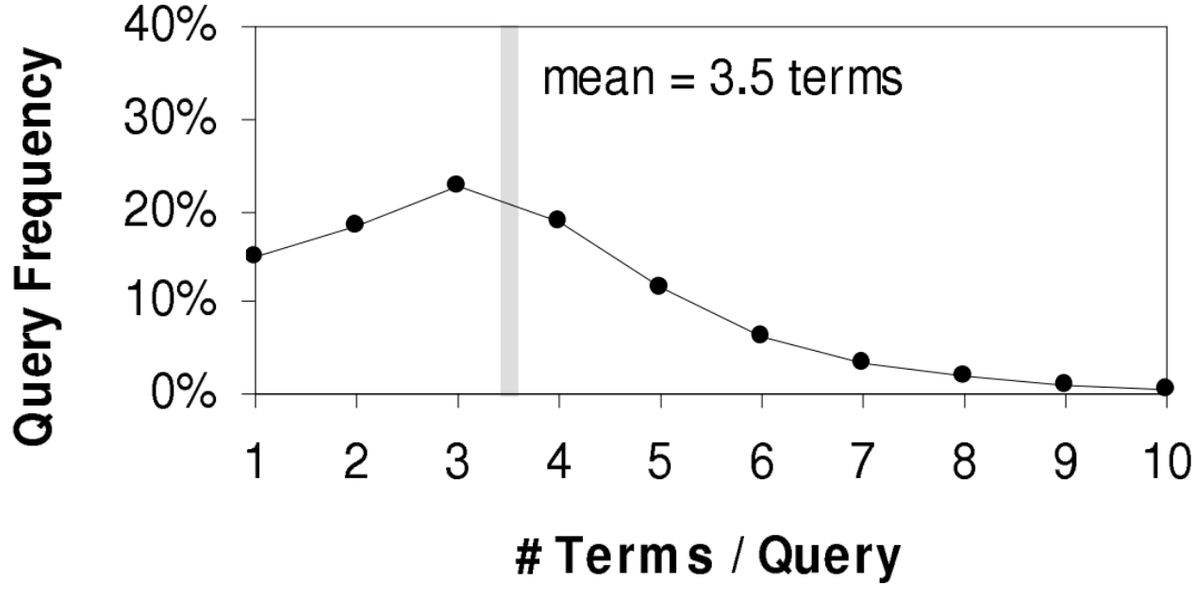
WHEN?

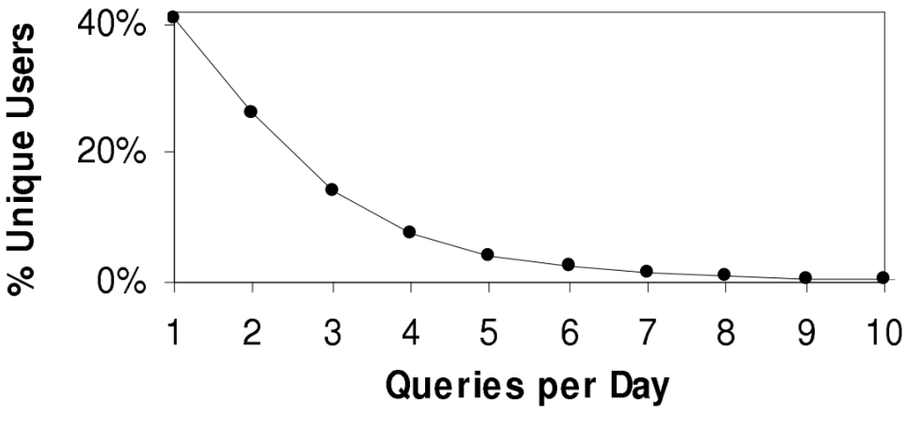


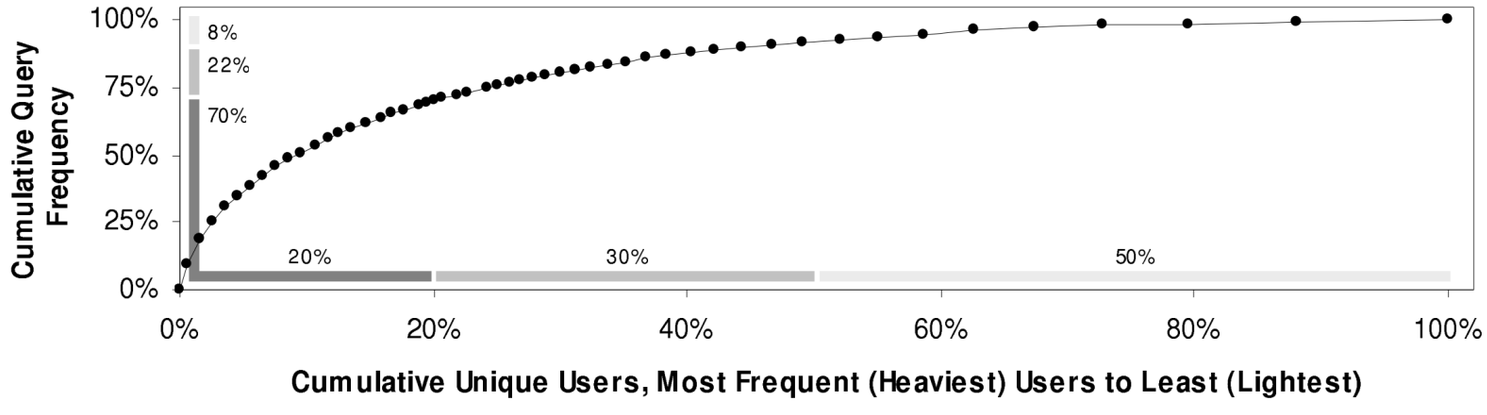
UNDERSTANDING SEARCHES/SEARCHERS

HOW? http://dl.acm.org/citation.cfm?id=1146848



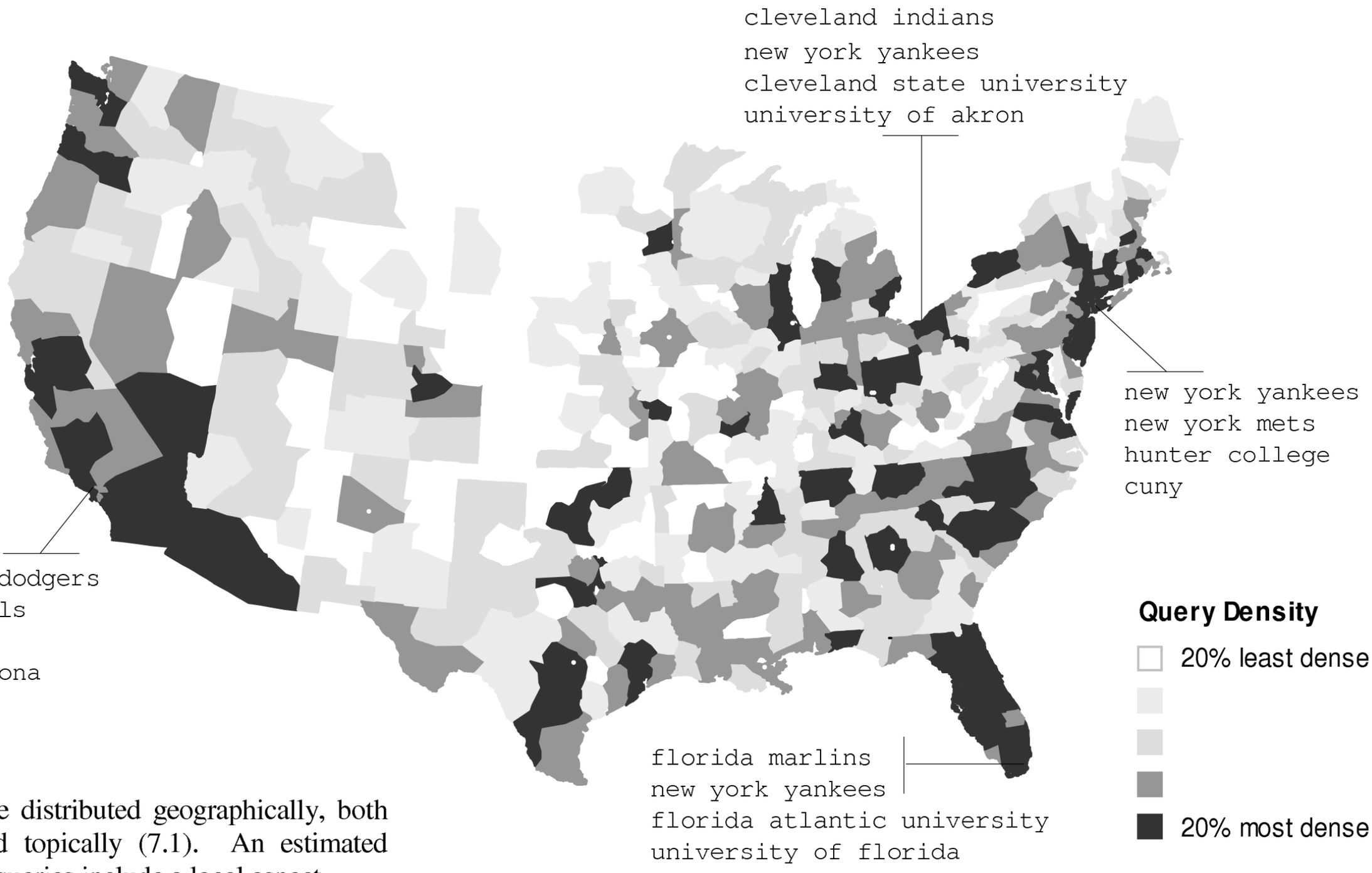






LECTURE 12 - SEARCH LOG ANALYSIS

(7.1)



User queries are distributed geographically, both numerically and topically (7.1). An estimated 12% to 28% of queries include a local aspect.

http://dl.acm.org/citation.cfm?id=1146848

Greg Pass & Abdur Chowdhury & Cayley Torgeson - InfoScale (2006)

MORE RECENT INFORMATION

- 20% of all queries seen each day have never been seen before (stats from Bing and Google)
- ~8% of all queries are names (Amit Singhal, Google, 2010)

TASKS

TASKS

UNDERSTANDING USERS

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

Ingmar Weber & Alejandro Jaimes WSDM 2011

- Same three dimensions that we just spoke about:
- What (topics):
  - Similarity checked from Yahoo! Directory, dmoz.com
- Who (user demographics):
  - Provided by the user (age, gender)
  - Inferred by user's zip code: income, education level, political party affiliation (Demography / Census / Pew Research / Vox Populi data)
- How (session characteristics):
  - Session length, variations in the document people click on

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Data from Yahoo! search engine (2008-2009)
  - Active users (> 100 queries during sample period)
    - U.S. users
    - 2.3 million users (~1% U.S. population)
- Cluster users based on the types of queries they issued

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**
- Pseudo document for each user:

AnonID	Query	Yahoo Category
142	rentdirect.com	Finance
142	staple.com	Office
142	dfdf	None/Removed
142	lottery	Betting Games
142	ameriprise.com	Finance
...	...	...

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**
- Pseudo document for each user:

AnonID	Query	Yahoo Category
142	rentdirect.com	Finance
142	staple.com	Office
142	dfdf	None/Removed
142	lottery	Betting Games
142	ameriprise.com	Finance
…	…	…

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**
- Pseudo document for each user:

AnonID	Query	Yahoo Category
142	rentdirect.com	Finance
142	staple.com	Office
142	dfdf	None/Removed
142	lottery	Betting Games
142	ameriprise.com	Finance
...	...	...

```
<DOC>
<TITLE> 142 </TITLE>
<TEXT>
FINANCE OFFICE BETTING GAMES
FINANCE ARCHITECTURE DESIGN
FINANCE
</TEXT>
</DOC>
```

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**
- Pseudo document for each user:
- Richer representation:

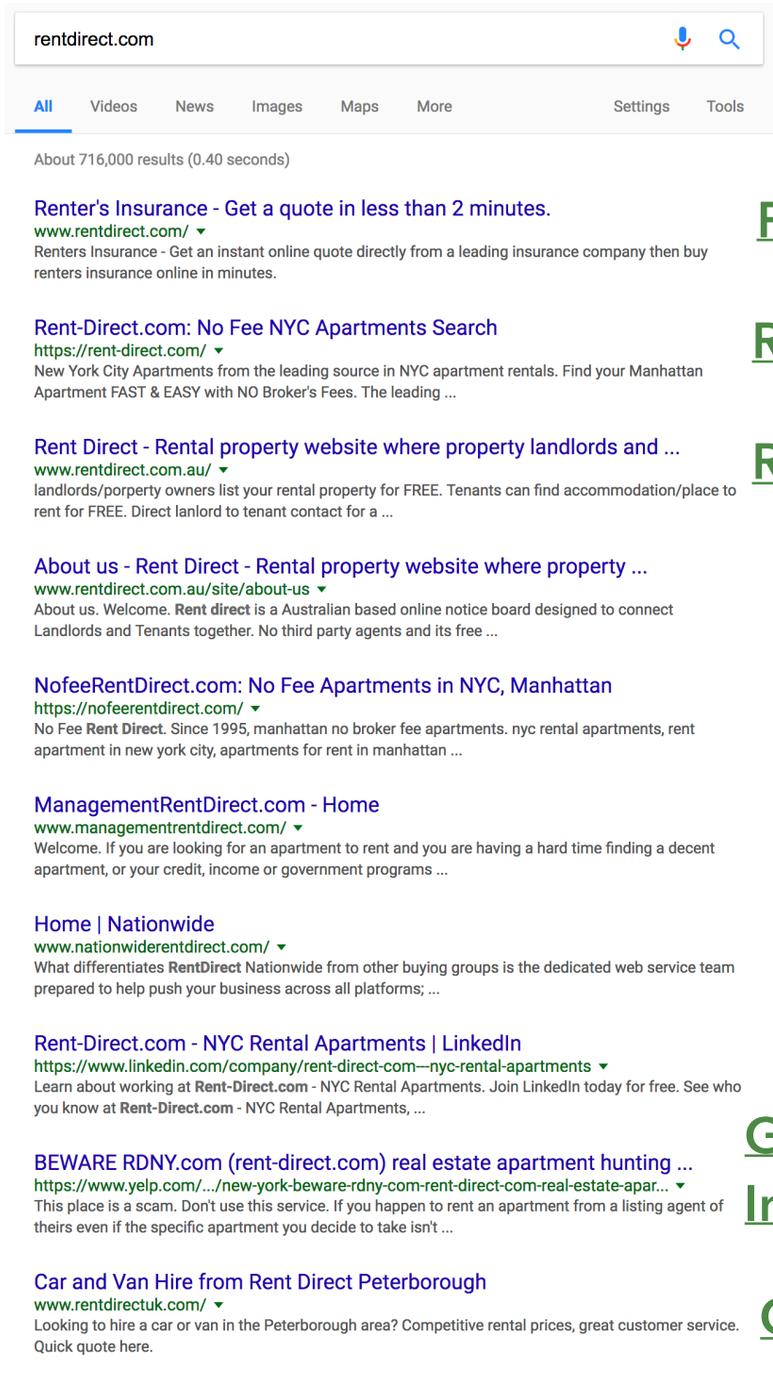
AnonID	Query
142	rentdirect.com
142	staple.com
142	dfdf
142	lottery
142	ameriprise.com
…	…

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Clustering users? **HOW?**
- Pseudo document for each user:
- Richer representation:

AnonID	Query
142	rentdirect.com
142	staple.com
142	dfdf
142	lottery
142	ameriprise.com
...	...

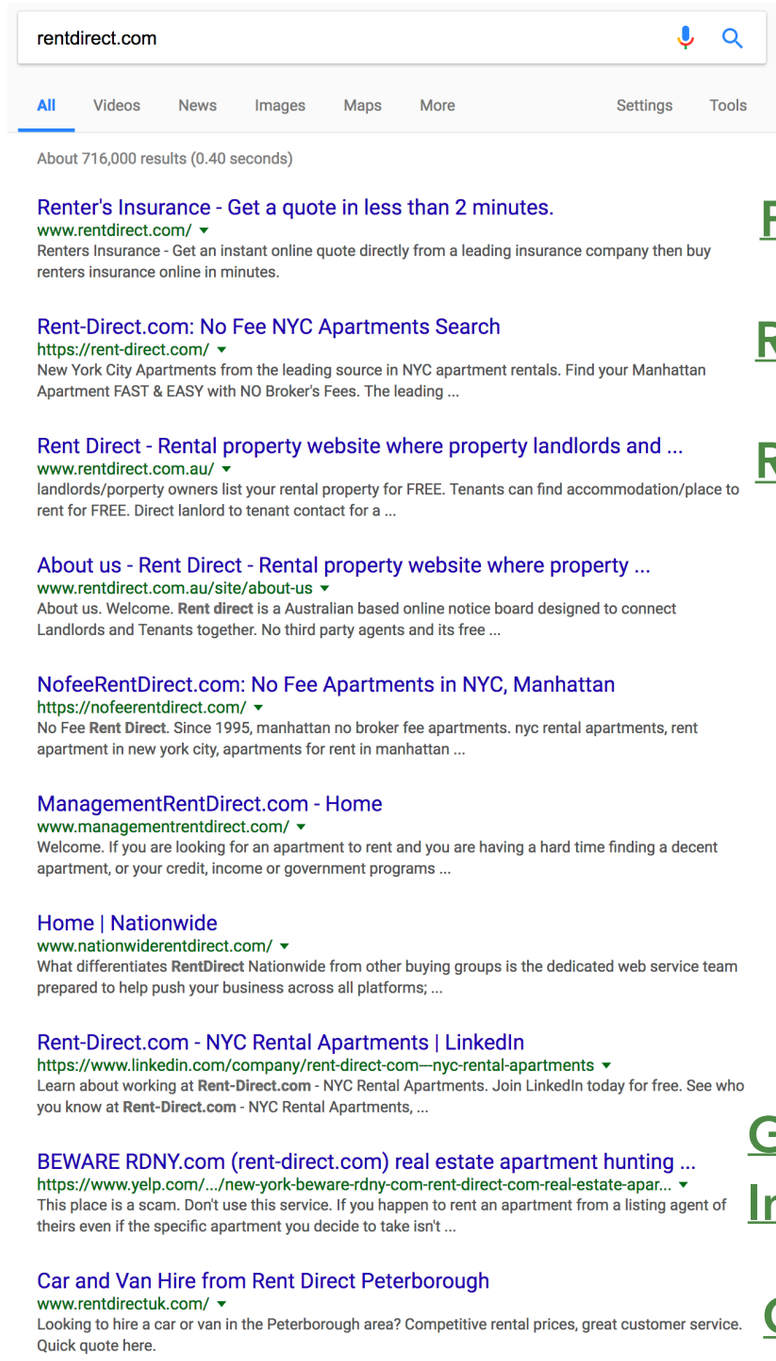


UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

```
<DOC>
<TITLE> 142 </TITLE>
<TEXT>
FINANCE REAL ESTATE REAL ESTATE
.... GENERAL INFORMATION CAR
RENTAL... BETTING GAMES FINANCE
ARCHITECTURE DESIGN FINANCE
</TEXT>
```

AnonID	Query
142	rentdirect.com
142	staple.com
142	dfdf
142	lottery
142	ameriprise.com
...	...



UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

<DOC>
<TITLE> 142 </TITLE>
<TEXT>
FINANCE REAL ESTATE REAL ESTATE GENERAL INFORMATION CAR RENTAL BETTING GAMES FINANCE ARCHITECTURE DESIGN FINANCE
</TEXT>

<DOC>
<TITLE> 6120607 </TITLE>
<TEXT>
RESTAURANTS CHINESE FOOD ADULT PORN CHURCH CHURCH EQUIPMENT PRIEST ASIAN PORN FACIAL PORN BIBLE LESSONS . . . </TEXT>
</DOC>

<DOC>
<TITLE> 3348270 </TITLE>
<TEXT>
ALCOHOL BIER DRINKS CAIPIRINHA VODKA AA ALCOHOLICS ANONYMOUS AMERICA AIRLINES
</TEXT>
</DOC>

- Use your favorite clustering algorithm with your favorite similarity measure
- Same idea is used by people search (Linkedin), product search and for personalization

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Users clustered in the “what” dimension
- Explore the “who” and “how” dimensions

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Users clustered in the “what” dimension
- Explore the “who” and “how” dimensions

- Clustering is unsupervised technique: No Labels!!!
  - Manually people looked at these clusters to characterize each of them

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Cluster 1: Navigational Users
- How:
  - Navigational queries (facebook, google, qnb)
  - Single-click sessions
  - Less likely to use query suggestions
- What:
  - Popular Websites
- Who:
  - Spread over the entire population

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Cluster 2: Informational Users
- How:
  - Non-Navigational queries (some one doing some kind of research)
  - Longer sessions
  - More likely to use query suggestions
- What:
  - Wide rage of topics (less adult content)
- Who:
  - More likely to be well-educated
  - More likely to have above-average income

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Cluster: Baby boomers:
  - 50 years old
  - Interested in Finance
  - Simple navigational queries related to online banking
- Cluster: Challenged Youth
  - Average age of 34
  - Low-income, low-level of education
  - Interested in music
  - Navigational queries

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Cluster: Liberal Females
  - Mostly female from areas that voted Democratic
  - Shopping queries
  - Long sessions (browsing and comparison)
- Cluster: White Conservatives
  - Mostly male from areas that voted Republican
  - Interest in automotive, business, home & garden

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Cluster: Liberal Females

NOTE THE HIGHLY STEREOTYPICAL ANALYSIS

  - Mostly female from areas that voted Democratic
  - Shopping queries
  - Long sessions (browsing and comparison)
- Cluster: White Conservatives
  - Mostly male from areas that voted Republican
  - Interest in automotive, business, home & garden

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Older users: health, diseases, gambling, travel
- Teenagers: Games, education
- Asian descent: Computers, internet, programming, software development
- Low income: music, comics, animation, military, games

UNDERSTANDING/PROFILING USERS

http://dl.acm.org/citation.cfm?id=1935839

- Older users: health, diseases, gambling, travel
- Teenagers: Games, education
- Asian descent: Computers, internet, programming, software development
- Low income: music, comics, animation, military, games

AGAIN: CONFIRMATION BIAS?

TASKS

SEGMENTING QUERY LOGS

WHAT IS A SEARCH SESSION?

- Correct identification of a search sessions is a key task
- All the other tasks we will see today assume we have a good way to identify sessions...

WHAT IS A SEARCH SESSION?

- Correct identification of a search sessions is a key task
- All the other tasks we will see today assume we have a good way to identify sessions…
- However, that is still an open research problem…

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

WHAT IS A SEARCH SESSION?

Query	QueryTime
gout	2006-03-01 07:38:03
chemotherapy side effects	2006-03-01 07:42:36
chemotherapy causing hearing loss	2006-03-01 07:45:23
kenny rogers songs	2006-03-02 06:05:40
commerce on line	2006-03-03 04:54:11
broadband internet	2006-03-06 05:32:28
middlesex county college nj	2006-03-06 16:55:56
kean college	2006-03-06 17:02:32
montclair college	2006-03-06 17:10:45
union county college	2006-03-07 04:49:23
rutgers	2006-03-07 05:10:17
kean college	2006-03-07 05:19:22
migrane headache	2006-03-10 06:02:55
new jersey income tax	2006-04-12 06:09:44

HEURISTICS

- Researchers have been relying on simple heuristics:
  - **Time:**
    - Same session iff: $|timestamp(q_2) - timestamp(q_1)| < T$
  - **Common terms:**
    - Same session iff $terms(q_1)$ intersection $terms(q_2) > 0$
  - **Rewrite classes:**
    - Same session iff term added, deleted or replaced

HEURISTICS

Query	QueryTime	
gout	2006-03-01 07:38:03	CT, RC
chemotherapy side effects	2006-03-01 07:42:36	
chemotherapy causing hearing loss	2006-03-01 07:45:23	Time,CT, RC
kenny rogers songs	2006-03-02 06:05:40	
commerce on line	2006-03-03 04:54:11	
broadband internet	2006-03-06 05:32:28	
middlesex county college nj	2006-03-06 16:55:56	
kean college	2006-03-06 17:02:32	
montclair college	2006-03-06 17:10:45	Time
union county college	2006-03-07 04:49:23	CT, RC
rutgers	2006-03-07 05:10:17	CT, RC
kean college	2006-03-07 05:19:22	Time,CT, RC
migrane headache	2006-03-10 06:02:55	Time,CT, RC
new jersey income tax	2006-04-12 06:09:44	

MAIN CHALLENGES

- Information need may span days, weeks, months...
  - Planning a trip, writing a paper, choosing a university

- People are multitask:
  - Can stop a search session to do another one
  - Can stop a search session to have lunch

- An information need might have multiple subtasks:
  - Some of them might appear distinct when they are actually very related

TYPICALLY SOLVED WITH ML

- Classification: Do these pair of (sequential) queries belong to the same information need?

TYPICALLY SOLVED WITH ML

- Classification: Do these pair of (sequential) queries belong to the same information need?
- Additional Features:
  - Edit distance between queries
  - Co-occurrence in a query log
- Overlap of page categories of top 10 results
- Average cosine distance of top 50 results

RESULTS

- Predict whether two queries are for the same information need:
  - Adjacent queries: 85-90% accuracy
  - Any pair of queries: 95-97% accuracy

RESULTS

- Predict whether two queries are for the same information need:
  - Adjacent queries: 85-90% accuracy
  - Any pair of queries: 95-97% accuracy

INTERPRETATION OF RESULTS:
WHY IS THE ACCURACY HIGHER FOR THIS TASK?

RESULTS

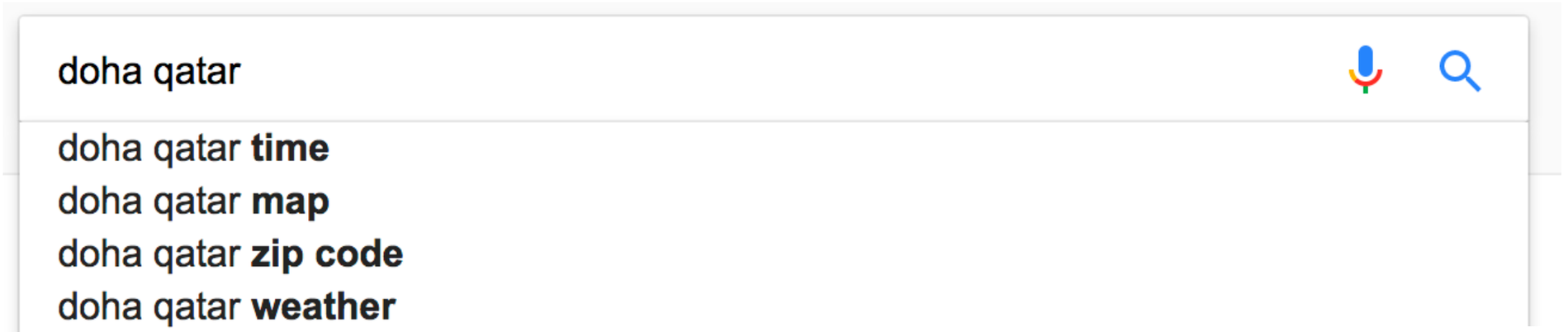
- Predict whether two queries are for the same information need:
  - Adjacent queries: 85-90% accuracy
  - Any pair of queries: 95-97% accuracy
- Main message:
  - Classifiers work best, but heuristics alone are not far behind:
    - Edit distance is very effective
    - Cosine distance of results is very effective
    - Even time thresholds are competitive baselines and often used.

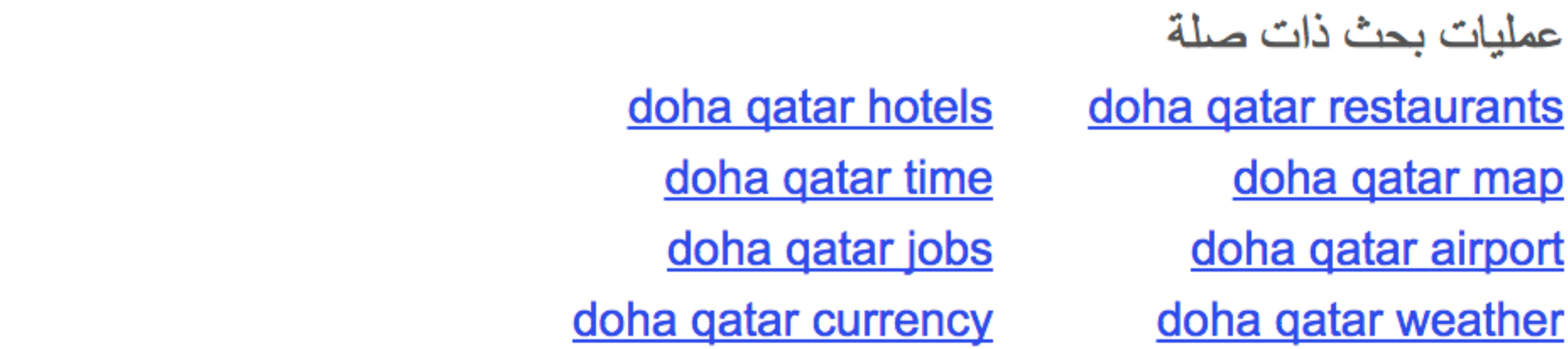
TASKS

QUERY SUGGESTIONS

QUERY SUGGESTIONS

- Task:
  - Given a query, suggest a completion or another query





QUERY SUGGESTIONS

- How can we generate query suggestions?

QUERY SUGGESTIONS

- How can we generate query suggestions?
- Simple idea:
  - Rank all the queries ever issued by a criteria such as the query popularity

QUERY SUGGESTIONS

- How can we generate query suggestions?
- Simple idea:
  - Rank all the queries ever issued by a criteria such as the query popularity
- Some problems:
  - 97% of queries occurs less than 10 times
  - 57% of queries are unique

QUERY SUGGESTIONS

- You should expect:
  - Can we explore search sessions to obtain query reformulations/suggestions?
- Methods:
  - Pseudo-documents
  - Co-occurrence

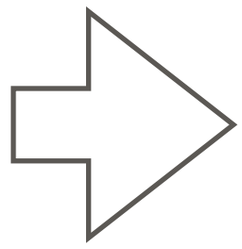
QUERY SUGGESTIONS – PSEUDO-DOCUMENTS

- Joao's favorite method: pseudo-documents!

QUERY SUGGESTIONS – PSEUDO-DOCUMENTS

- Joao's favorite method: pseudo-documents!
- Idea obtain from query logs pairs like <$query_i$, $query_{last}$>:

From AOL logs:
1185719, housing in doha qatar
1185719, villa apartment doha qatar
1185719, residential villas in doha qatar
1185719, real estate in doha qatar

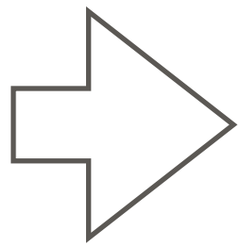


<housing in doha qatar, real estate in doha qatar>
< villa apartment doha qatar, real estate in doha qatar>
< residential villas in doha qatar, real estate in doha qatar>
< [from another session], real estate in doha qatar>

QUERY SUGGESTIONS – PSEUDO-DOCUMENTS

- Joao's favorite method: pseudo-documents!
- Idea obtain from query logs pairs like <$query_i$, $query_{last}$>:

From AOL logs:
1185719, housing in doha qatar
1185719, villa apartment doha qatar
1185719, residential villas in doha qatar
1185719, real estate in doha qatar

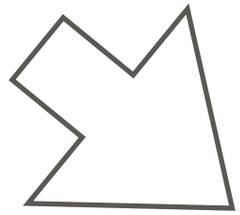


<housing in doha qatar, real estate in doha qatar>
< villa apartment doha qatar, real estate in doha qatar>
< residential villas in doha qatar, real estate in doha qatar>
< [from another session] , real estate in doha qatar>

- Strong assumption: last query in a session is successful

QUERY SUGGESTIONS – PSEUDO-DOCUMENTS

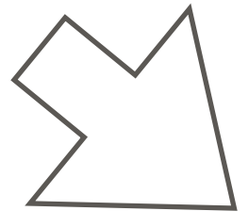
<housing in doha qatar, real estate in doha qatar>
< villa apartment doha qatar, real estate in doha qatar>
< residential villas in doha qatar, real estate in doha qatar>



<DOC>
<TITLE> REAL ESTATE IN DOHA QATAR </TITLE>
<TEXT>
HOUSING IN DOHA QATAR VILLA APARTMENT DOHA QATAR RESIDENTIAL VILLAS IN DOHA QATAR REAL
</TEXT>
</DOC>

QUERY SUGGESTIONS – PSEUDO-DOCUMENTS

<housing in doha qatar, real estate in doha qatar>
< villa apartment doha qatar, real estate in doha qatar>
< residential villas in doha qatar, real estate in doha qatar>



<DOC>
<TITLE> REAL ESTATE IN DOHA QATAR </TITLE>
<TEXT>
HOUSING IN DOHA QATAR VILLA APARTMENT DOHA QATAR RESIDENTIAL VILLAS IN DOHA QATAR REAL
</TEXT>
</DOC>

- Given a new query (even one that was never seen before):
  1. Apply your favorite retrieval model to the pseudo document corpus
  2. Suggest top X query titles for a user

QUERY SUGGESTIONS - CO-OCCURRENCES

- Given a query q:
  - Set 1: 100 most frequent queries with q as a substring
  - Set 2: 100 most frequent queries that followed q in a search session

QUERY SUGGESTIONS – CO-OCCURRENCES

- Given a query q:
  - Set 1: 100 most frequent queries with q as a substring
  - Set 2: 100 most frequent queries that followed q in a search session
- SuperSet: Set 1 + Set 2
- For each q_s in SuperSet:

$$Score(q_s) = \frac{Count(q_s) + \lambda_1}{N_1 + \lambda_1} \times \frac{Count_{follows}(q, q_s) + \lambda_2}{N_2 + \lambda_2}$$

QUERY SUGGESTIONS – CO-OCCURRENCES

- Given a query q:
  - Set 1: 100 most frequent queries with q as a substring
  - Set 2: 100 most frequent queries that followed q in a search session
- SuperSet: Set 1 + Set 2
- For each q_s in SuperSet:

$$Score(q_s) = \frac{Count(q_s) + \lambda_1}{N_1 + \lambda_1} \times \frac{Count_{follows}(q, q_s) + \lambda_2}{N_2 + \lambda_2}$$

q_s is frequent

q_s frequently follows q

Smoothing Parameters

QUERY SUGGESTIONS – CO-OCCURRENCES

- Given a query q:
  - Set 1: 100 most frequent queries with q as a substring
  - Set 2: 100 most frequent queries that followed q in a search session
- SuperSet: Set 1 + Set 2
- For each q_s in SuperSet:

$$Score(q_s) = \frac{Count(q_s) + \lambda_1}{N_1 + \lambda_1} \times \frac{Count_{follows}(q, q_s) + \lambda_2}{N_2 + \lambda_2}$$

q_s is frequent

q_s frequently follows q

Smoothing Parameters

http://dl.acm.org/citation.cfm?id=1135835

QUERY SUGGESTIONS – CO-OCCURRENCES #2

- Is it more likely to observe a pair of queries together or independently?
- For 2 queries in the same session q_1 and q_2:
  - H1: Seen q_2 is equally likely whether or not we have seen q_1
  - H2: Seen q_2 is more / less likely if we have seen q_1

Paper uses a log likelihood ratio test to rank the candidates:

$$\mathrm{LLR} = -2\log\lambda = -2\log\frac{L(H_1)}{L(H_2)}$$

http://dl.acm.org/citation.cfm?id=1135835

QUERY SUGGESTIONS - CO-OCCURRENCES #2

dog → dogs	9185	(pluralization)
dog → cat	5942	(both instances of 'pet')
dog → dog breeds	5567	(generalization)
dog → dog pictures	5292	(more specific)
dog → 80	2420	(random junk or noise)
dog → pets	1719	(generalization – hypernym)
dog → puppy	1553	(specification – hyponym)
dog → dog picture	1416	(more specific)
dog → animals	1363	(generalization – hypernym)
dog → pet	920	(generalization – hypernym)

Table 2: Terms and queries which can be substituted for the term or query "dog", along with likelihood ratios, based on user query rewriting sessions. The semantic relationship is shown for explanatory purposes only.

TASKS

QUERY INTENTS

AMBIGUITY

- We know ambiguity is a problem:
  - jaguar: a car, an animal, an operating system, …
  - flash: software, a superhero, part of a camera…
  - mercury: an element, a planet, a car, a hotel, a god
  - ai: artificial intelligence, american idol, air india
  - michael jordan: an athlete, a professor, a businessman
- Diversity in the search results is a widely accepted solution

QUERY INTENTS

- If a query is very ambiguous might and you do not know anything about the user, the best idea is to suggest the different meanings (maybe ranked by their likelihood)

QUERY INTENTS

- If a query is very ambiguous might and you do not know anything about the user, the best idea is to suggest the different meanings (maybe ranked by their likelihood)

A.I. Artificial Intelligence (2001) - **IMDb** — Movie
www.**imdb**.com/title/tt0212720
7.1/10 · 249K ratings · Adventure/Drama/Sci-Fi · PG-13
Directed by Steven Spielberg. With Haley Joel Osment, Jude Law, Frances O'Connor, Sam Robards. A highly advanced robotic boy longs to become "real" so that he can ...

Adobe Illustrator Artwork - Wikipedia — Software
https://**en.wikipedia.org**/wiki/**Adobe_Illustrator_Artwork**
Adobe Illustrator Artwork (**AI**) is a proprietary file format developed by Adobe Systems for representing single-page vector-based drawings in either the EPS or PDF ...

Ai | Define **Ai** at **Dictionary**.com — Dictionary
www.**dictionary**.com/browse/**ai**
Ai definition, a three-toed sloth, Bradypus tridactylus, inhabiting forests of southern Venezuela, the Guianas, and northern Brazil, having a diet apparently ...

AI - definition of **AI** by The Free **Dictionary**
www.thefreedictionary.com/AI
AI abbr. 1. airborne intercept 2. **Amnesty International** 3. aromatase inhibitor 4. **artificial insemination** 5. **artificial intelligence ai** (ī) n. See sloth. [Portuguese ...

AI File Extension - What is a .**ai** file and how do I open it? — File extension
https://fileinfo.com/extension/**ai**
3.8/5 · 144 ratings
An **AI** file is a drawing created with **Adobe Illustrator**, a vector graphics editing program. It is composed of paths connected by points, rather than bitmap image data.

Ai Research - Creating a new form of life — Company
www.**a-i**.com
Artificial Intelligence Ltd. (**Ai**) develops conversational software - technology that enables machines to converse with humans in natural language.

AI Official Site : LIVE — Music
aimusic.tv/live
AI MUSIC; VIDEOS; **MUSIC**; PHOTOS; ABOUT; xx close xx. NEWS; LIVE; STORE; **AI MUSIC**; VIDEOS; **MUSIC**; PHOTOS; ABOUT; Worldwide booking and press (except …

A.I. Experiments — Google
https://**aiexperiments**.withgoogle.com
AI Experiments is a showcase for simple experiments that let anyone play with **artificial intelligence** and machine learning in hands-on ways, through pictures ...

Music - YouTube — WTF?
www.youtube.com/channel/UC-9-kyTW8ZkZNDHQJ6FgpwQ
YouTube's **music** destination featuring top tracks and popular hits from a variety of genres. This channel was generated automatically by **YouTube**'s video disco...

Air India Limited - Official Site — Air India
www.**airindia**.com
Official website of the country's national carrier. Includes flight schedules, reservations, in flight services frequent flyer plans and cargo information.

QUERY INTENTS

MISSING: WE NEED TO KNOW THE POTENTIAL QUERY INTENTS!!!

- If a query is very ambiguous might and you do not know anything about the user, the best idea is to suggest the different meanings (maybe ranked by their likelihood)

A.I. Artificial Intelligence (2001) - **IMDb**
www.imdb.com/title/tt0212720
★★★½☆ **7.1/10** · 249K ratings · Adventure/Drama/Sci-Fi · PG-13
Directed by Steven Spielberg. With Haley Joel Osment, Jude Law, Frances O'Connor, Sam Robards. A highly advanced robotic boy longs to become "real" so that he can ...

Movie

Adobe Illustrator Artwork - Wikipedia
https://**en.wikipedia.org**/wiki/**Adobe_Illustrator_Artwork**
Adobe Illustrator Artwork (**AI**) is a proprietary file format developed by Adobe Systems for representing single-page vector-based drawings in either the EPS or PDF ...

Software

Ai | Define **Ai** at **Dictionary**.com
www.**dictionary**.com/browse/**ai**
Ai definition, a three-toed sloth, Bradypus tridactylus, inhabiting forests of southern Venezuela, the Guianas, and northern Brazil, having a diet apparently ...

Dictionary

AI - definition of **AI** by The Free **Dictionary**
www.thefreedictionary.com/AI
AI abbr. 1. airborne intercept 2. **Amnesty International** 3. aromatase inhibitor 4. **artificial insemination** 5. **artificial intelligence ai** (ī) n. See sloth. [Portuguese ...

AI File Extension - What is a .**ai** file and how do I open it?
https://fileinfo.com/extension/**ai**
★★★★☆ **3.8/5** · 144 ratings
An **AI** file is a drawing created with **Adobe Illustrator**, a vector graphics editing program. It is composed of paths connected by points, rather than bitmap image data.

File extension

Ai Research - Creating a new form of life
www.**a-i**.com
Artificial Intelligence Ltd. (**Ai**) develops conversational software - technology that enables machines to converse with humans in natural language.

Company

AI Official Site : LIVE
aimusic.tv/live
AI MUSIC; VIDEOS; **MUSIC**; PHOTOS; ABOUT; xx close xx. NEWS; LIVE; STORE; **AI MUSIC**; VIDEOS; **MUSIC**; PHOTOS; ABOUT; Worldwide booking and press (except …

Music

A.I. Experiments
https://**aiexperiments**.withgoogle.com
AI Experiments is a showcase for simple experiments that let anyone play with **artificial intelligence** and machine learning in hands-on ways, through pictures ...

Google

Music - YouTube
www.youtube.com/channel/UC-9-kyTW8ZkZNDHQJ6FgpwQ
YouTube's **music** destination featuring top tracks and popular hits from a variety of genres. This channel was generated automatically by **YouTube**'s video disco...

WTF?

Air India Limited - Official Site
www.**airindia**.com
Official website of the country's national carrier. Includes flight schedules, reservations, in flight services frequent flyer plans and cargo information.

Air India

http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)

http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
  - Identify the 10 most common reformulations q'' of q'

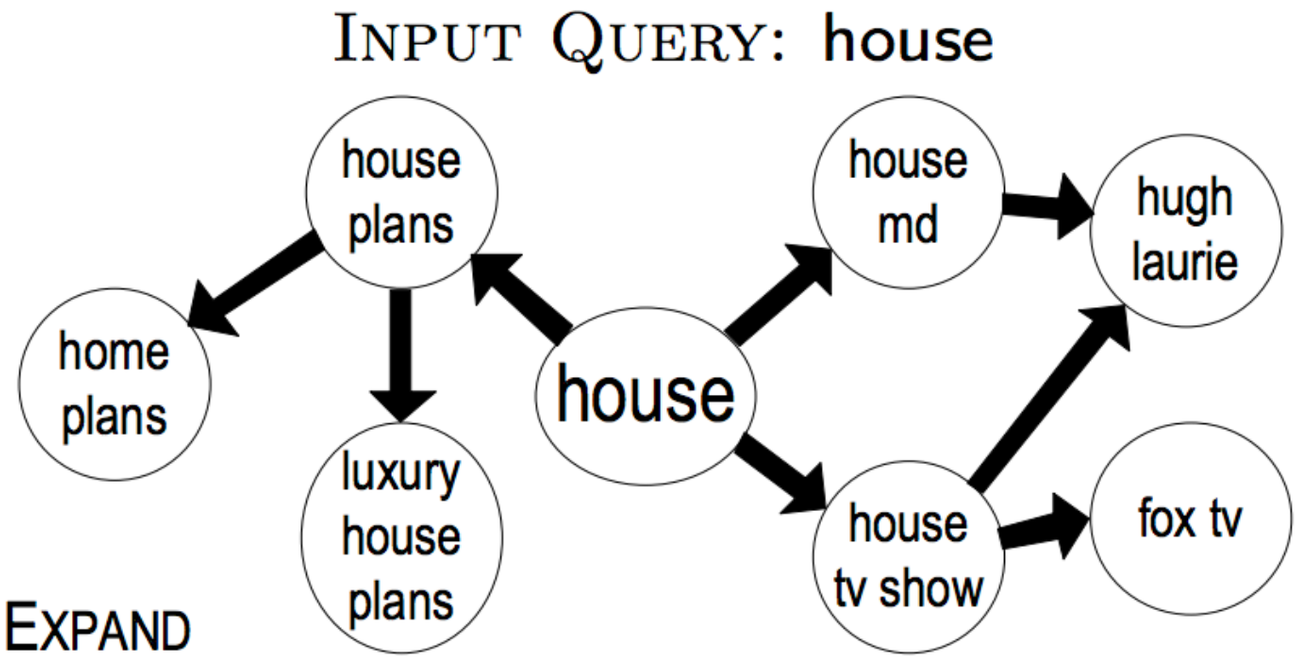
http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
  - Identify the 10 most common reformulations q'' of q'

Connect Q1 and Q2 if:

At least Y users issued sequence
TimeWindow (Q1, Q2) < X min
SessionsWith(Q1,Q2) > Z



http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)

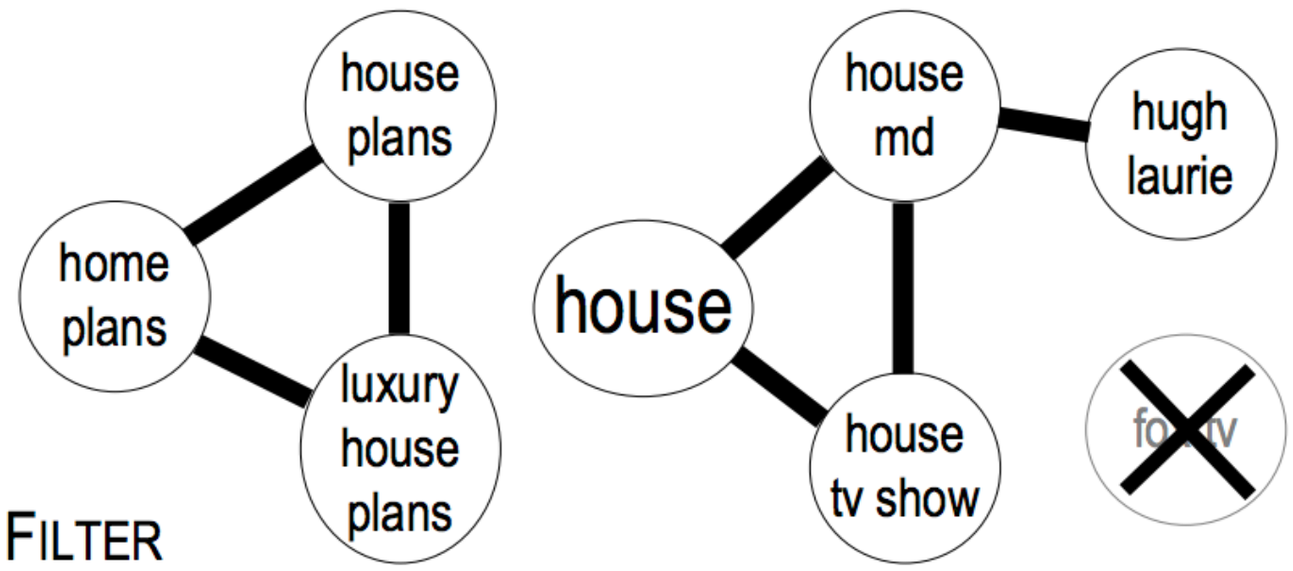
http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)

Remove link Q1 to Q2 if:

Retrieved set of results is not similar
Users do not click on the same documents



http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

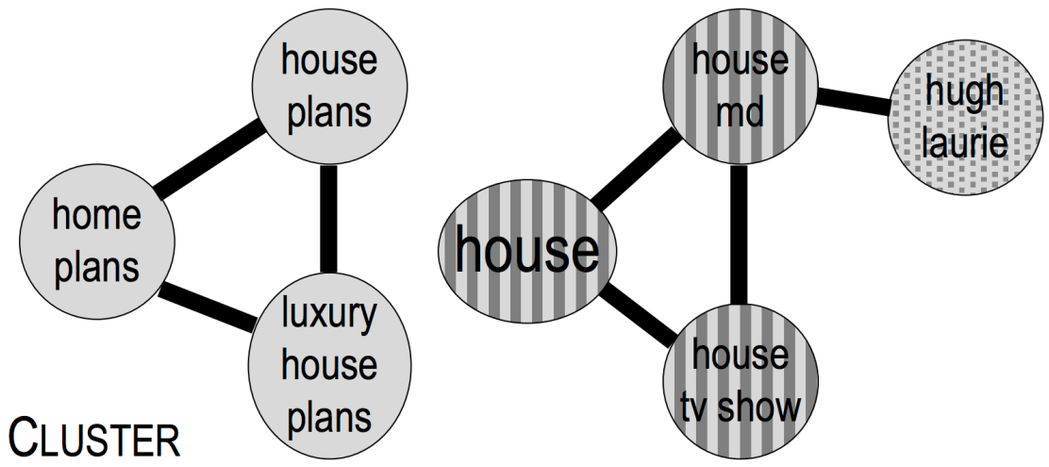
- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)
- Cluster results to find intent groups (cluster phase)

http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)
- Cluster results to find intent groups (cluster phase)

Use your favorite clustering method



http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)
- Cluster results to find intent groups (cluster phase)
- Estimate popularity of each query in a group (estimation phase)

http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- Idea: explore query suggestions (that explore search logs)
- Given a query q: (expansion phase)
  - Identify the 10 most common reformulations q' of q
    - Identify the 10 most common reformulations q'' of q'
- Prune graph (filter phase)
- Cluster results to find intent groups (cluster phase)
- Estimate popularity of each query in a group (estimation phase)

Random walk the graph: PageRank

http://dl.acm.org/citation.cfm?doid=1772690.1772859

QUERY INTENTS

- For query like “house”:
  - Search engine issues 3 queries:
    - “house plans”
    - “house md”
    - “hugh laurie”
- Merge the results by a criteria:
  - Popularity?

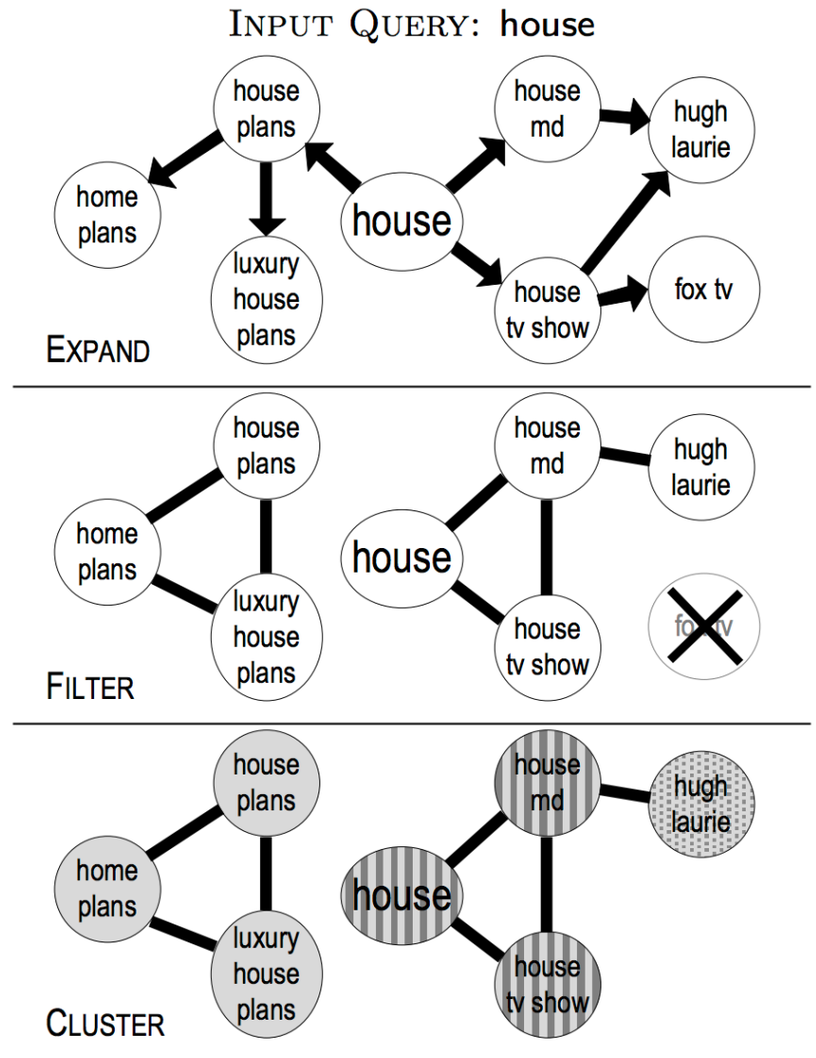


Figure 1: Three Steps of Intent Identification

Google

house

All | Images | Maps | News | More | Settings | Tools

About 5,760,000,000 results (0.73 seconds)

House
https://www.house.com.au/
House is Australia's largest independent and dedicated retailer of quality Homeware & Kitchenware products. **House** operates over 100 retail stores Australia ...
Store Locator · Specials · Kitchen · Kitchenware

The United States House of Representatives · House.gov
www.house.gov/
Home page of the United States **House** of Representatives.

House (TV series) - Wikipedia
https://en.wikipedia.org/wiki/House_(TV_series)
House (also called **House**, M.D.) is an American television medical drama that originally ran on the Fox network for eight seasons, from November 16, 2004 to ...
List of House episodes · Gregory House · Lisa Edelstein · Robert Sean Leonard

Images for house

More images for house | Report images

House - Wikipedia
https://en.wikipedia.org/wiki/House
A **house** is a building that functions as a home, ranging from simple dwellings such as rudimentary huts of nomadic tribes and the improvised shacks in ...

House of Fraser - Gifts, Fashion, Beauty, Home & Garden
https://www.houseoffraser.co.uk/
Shop from our range of women's, men's and kids' fashion, beauty, home, electricals and more at the UK's premium department store.

House Shop Online
www.housebrand.com/
New collection available now. Go to online catalog and discover more of newest fashion trends.

More of House Husbands - Watch | 9Now
https://www.9now.com.au/house-husbands
A unique drama that follows the chaotic lives of four families with one thing in common: the dads are in charge of raising the kids.

house.com.pl
www.house.pl/pl/en/
In 2017, give your wardrobe a makeover with **House**. You are young, adventurous, creative and inventive – show this off to the world! Be bold, express yourself ...

House M.D. (TV Series 2004–2012) - IMDb
www.imdb.com/title/tt0412142/
Drama · An antisocial maverick doctor who specializes in diagnostic medicine does whatever it takes to solve puzzling cases that come his way using his crack ...

house

Web | Images | News

31,600,000 RESULTS | Date | Language | Region

House M.D. (TV Series 2004–2012) - **IMDb**
www.imdb.com/title/tt0412142
8.8/10 · 326K ratings · Drama/Mystery · TV-14
A bus that **House** was riding crashes. **House** claims there's a victim on the bus that's dying, but not from the bus accident. He stops at nothing to figure out who the ...

United States House of Representatives - Official Site
www.house.gov
Home page of the **United States House of Representatives**

The House - Outdoor Gear, Outerwear & Bikes - Save up to …
www.the-house.com
A simple way to share the love. **The House** gift card personalizes your gift beyond the plastic. Give the gift that never needs a receipt.

Images of **house**
bing.com/images

See more images of **house**

House (1985) - **IMDb**
www.imdb.com/title/tt0091223
6.2/10 · 18K ratings · Comedy · R
Directed by Steve Miner. With William Katt, Kay Lenz, George Wendt, Richard Moll. A troubled writer moves into a haunted **house** after inheriting it from his aunt.

Houzz - Official Site
www.houzz.com
Houzz is the new way to design your home. Browse 13 million interior design photos, home decor, decorating ideas and home professionals online.

House TV Show: News, Videos, Full Episodes and More ...
www.tvguide.com/tvshows/**house**/100213
Watch full episodes of **House** and get the latest breaking news, exclusive videos and pictures, episode recaps and much more at **TVGuide.com**

House - Home | **Facebook**
www.facebook.com/House
HOUSE is a series in which the villain is a medical malady and the hero is a controversial doctor who trusts no one, least of all his patients.

house - Dizionario inglese-italiano WordReference
www.wordreference.com/enit/**house**
Compound Forms/Forme composte: Inglese: Italiano: acid **house** n noun: Refers to person, place, thing, quality, etc. ('80s, '90s dance music style) (stile musicale)

Google Images
images.google.com
Google Images. The most comprehensive image search on the web.

CONCLUSION

http://dl.acm.org/citation.cfm?id=1148175

MISSING TO COVER

- Click data:
  - Study where people click at:
    - Search Engine bias?
    - Navigational queries?
  - Interesting: user studies have shown that people often believe that first results are more trustable than others

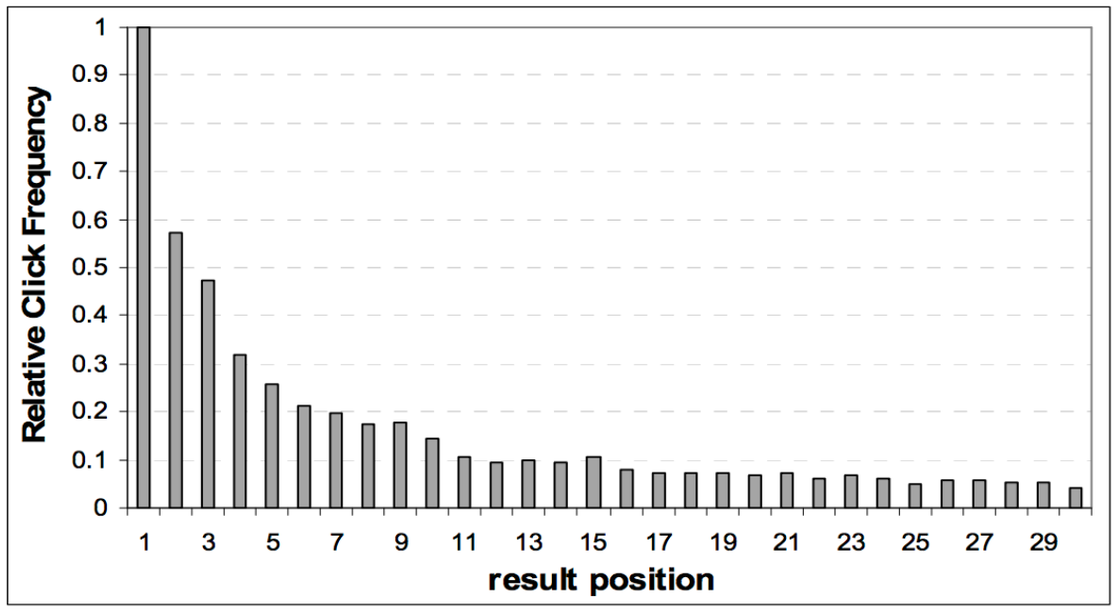


Figure 3.1: Relative click frequency for top 30 result positions over 3,500 queries and 120,000 searches.

http://dl.acm.org/citation.cfm?id=1148175

MISSING TO COVER

- Click data:
  - Study where people click at:
    - Search Engine bias?
    - Navigational queries?
    - Interesting: user studies have shown that people often believe that first results are more trustable than others
- Study why people do not click at results:
  - 50% of queries have no clicks!

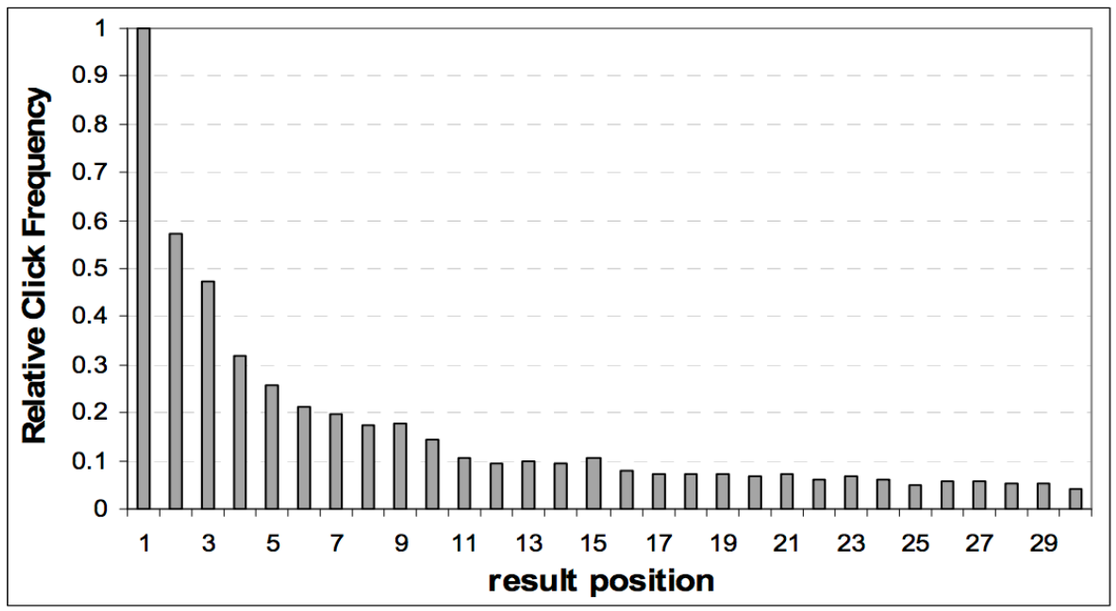


Figure 3.1: Relative click frequency for top 30 result positions over 3,500 queries and 120,000 searches.

IS IT A GOOD OR A BAD SIGNAL?

MISSING TO COVER

- Search logs can be used to many other tasks:
  - Cache policy:
    - Which queries to cache, how to cache, when to cache?
  - Teach user about a subject:
    - Detect if user is learning about a subject, increasingly expand user knowledge of the subject
  - Very good reference:
    - Mining Query Logs: Turning Search Usage Data into Knowledge

http://dl.acm.org/citation.cfm?id=1795387

WHAT DID WE SEE? WHAT SHOULD YOU KNOW?

- Most of the power of search engines come from **our inputs**
- Many interesting tasks built one on the top of the other: (1) understanding user behavior, (2) segmenting search logs into sessions, (3) query suggestions, (4) query intents

WHAT DID WE SEE? WHAT SHOULD YOU KNOW?

- Most of the power of search engines come from **our inputs**
- Many interesting tasks built one on the top of the other: (1) understanding user behavior, (2) segmenting search logs into sessions, (3) query suggestions, (4) query intents
- Techniques are simple:
  - Most important skill: creativity

WHAT DID WE SEE? WHAT SHOULD YOU KNOW?

- Most of the power of search engines come from **our inputs**
- Many interesting tasks built one on the top of the other: (1) understanding user behavior, (2) segmenting search logs into sessions, (3) query suggestions, (4) query intents
- Techniques are simple:
  - Most important skill: creativity
  - Turn a hard problem into something we know how to solve:
    - Vast use of **pseudo-documents**