Real Life, Real Users, and Real Needs : A Study and Analysis of User Queries on the Web

Understanding the Relationship between Searchers' Queries and Information Goals

Sung-Ju Fan-Chiang

2/19/2015

CS 597 Information Retrieval

Boise State University

Introduction

Internet IR vs. Traditional IR

Real life internet is changing IR ?

Internet IR is different IR ?

Trends ?

Trends in searching query

Search term length for business

User Behavior for Rare Versus Common Queries and Goals

Introduction

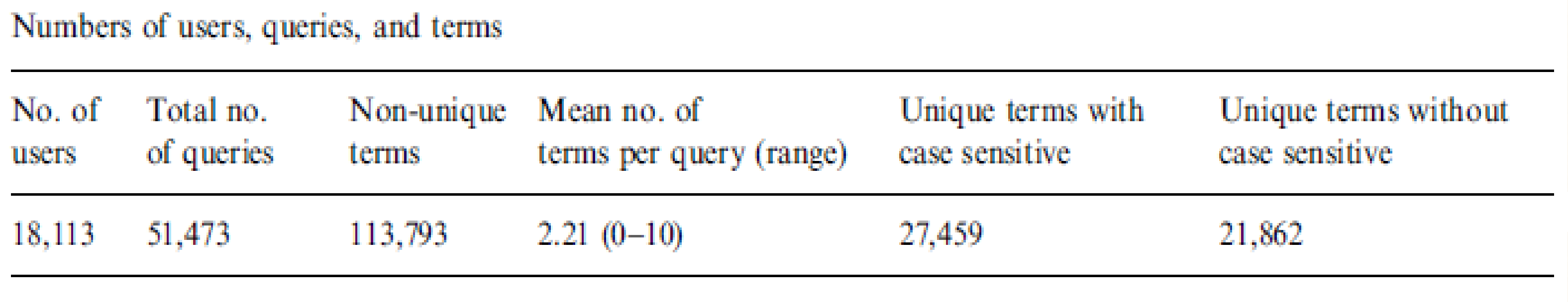
Background on **Excite** and data

Founded in 1994, Excite searchers are based on the exact terms that a user enters in the query.

Time of Day: measured in hours, minutes, and seconds from midnight of 9 March 1997.

User Identification: an anonymous user code assigned by the Excite server.

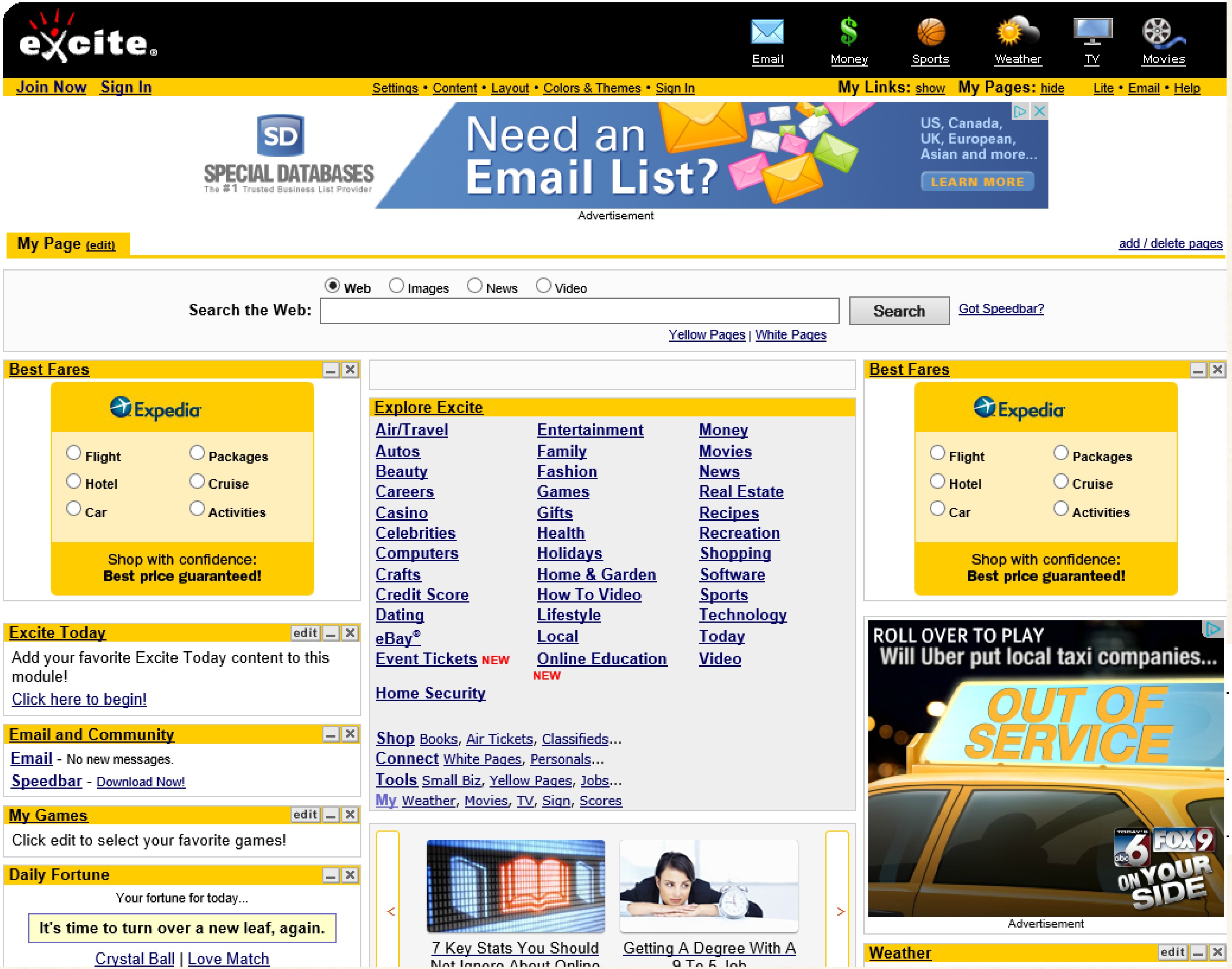
Query Terms: exactly as entered by the given user.



Numbers of users, queries, and terms

No. of users	Total no. of queries	Non-unique terms	Mean no. of terms per query (range)	Unique terms with case sensitive	Unique terms without case sensitive
18,113	51,473	113,793	2.21 (0–10)	27,459	21,862

Introduction



excite

Email | Money | Sports | Weather | TV | Movies

Join Now Sign In

Settings • Content • Layout • Colors & Themes • Sign In

My Links: show My Pages: hide Lite • Email • Help

My Page (edit)

add / delete pages

● Web ○ Images ○ News ○ Video

Search the Web: [ ] Search Got Speedbar?

Yellow Pages | White Pages

Best Fares

Expedia

○ Flight ○ Packages
○ Hotel ○ Cruise
○ Car ○ Activities

Shop with confidence:
Best price guaranteed!

Excite Today edit

Add your favorite Excite Today content to this module!

Click here to begin!

Email and Community

Email - No new messages.

Speedbar - Download Now!

My Games edit

Click edit to select your favorite games!

Daily Fortune

Your fortune for today...

It's time to turn over a new leaf, again.

Crystal Ball | Love Match

Explore Excite

Air/Travel
Autos
Beauty
Careers
Casino
Celebrities
Computers
Crafts
Credit Score
Dating
eBay®
Event Tickets NEW
Home Security

Entertainment
Family
Fashion
Games
Gifts
Health
Holidays
Home & Garden
How To Video
Lifestyle
Local
Online Education NEW

Money
Movies
News
Real Estate
Recipes
Recreation
Shopping
Software
Sports
Technology
Today
Video

Shop Books, Air Tickets, Classifieds...
Connect White Pages, Personals...
Tools Small Biz, Yellow Pages, Jobs...
My Weather, Movies, TV, Sign, Scores

<

7 Key Stats You Should Not Ignore About Online

Getting A Degree With A 9 To 5 Job

>

Best Fares

Expedia

○ Flight ○ Packages
○ Hotel ○ Cruise
○ Car ○ Activities

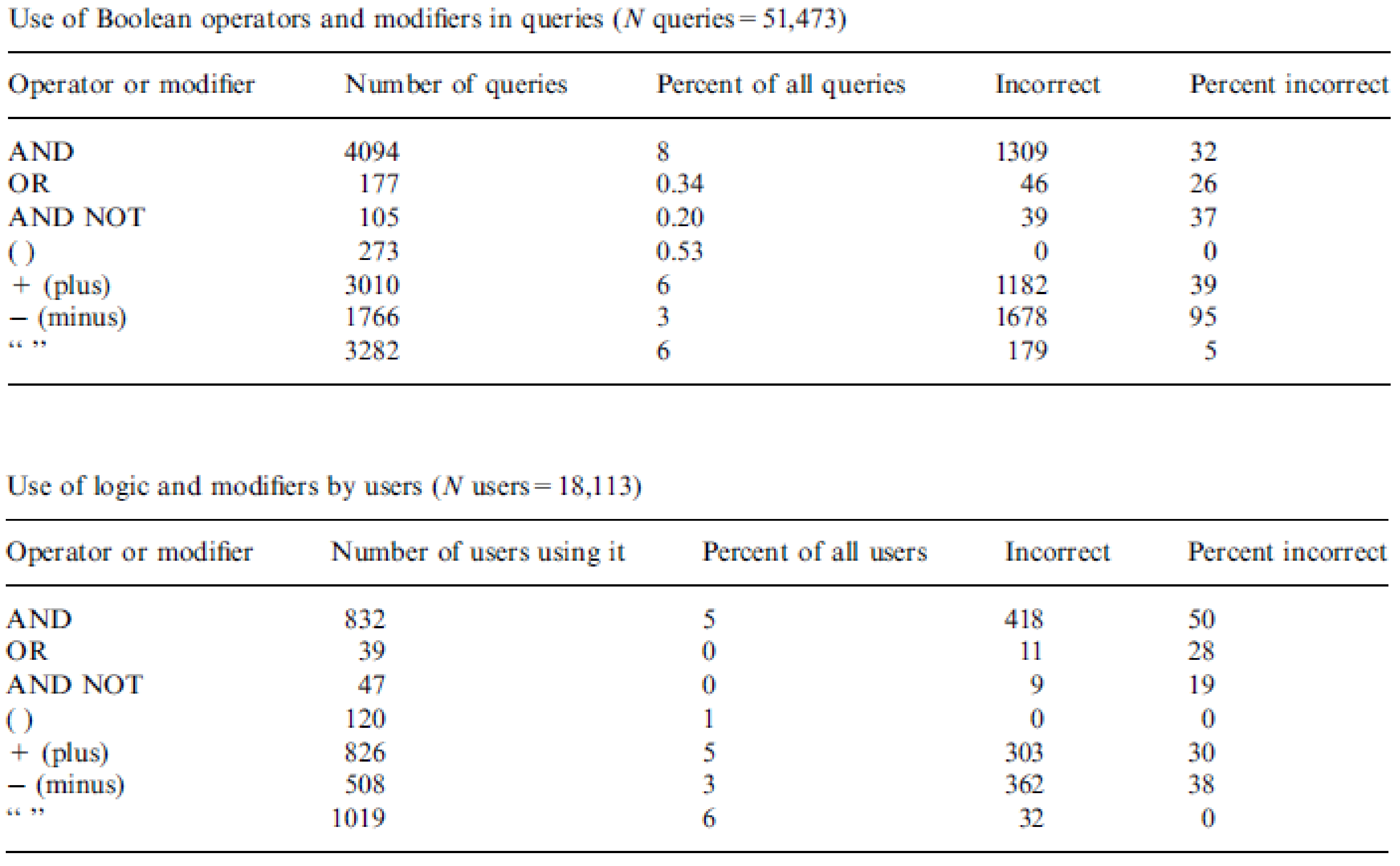
Shop with confidence:
Best price guaranteed!

Weather edit

Session, Query, Term, and Boolean

1. **Session**: A session is the entire series of queries by a user over a number of minutes or hours. A session could be as short as one query or contain many queries.
2. **Query**: A query consists of one or more search terms, and possibly includes logical operators and modifiers.
3. **Term**: A term is any unbroken string of characters (i.e. a series of characters with no space between any of the characters).

Session, Query, Term, and Boolean



Use of Boolean operators and modifiers in queries (*N* queries = 51,473)

Operator or modifier	Number of queries	Percent of all queries	Incorrect	Percent incorrect
AND	4094	8	1309	32
OR	177	0.34	46	26
AND NOT	105	0.20	39	37
( )	273	0.53	0	0
+ (plus)	3010	6	1182	39
– (minus)	1766	3	1678	95
“ ”	3282	6	179	5

Use of logic and modifiers by users (*N* users = 18,113)

Operator or modifier	Number of users using it	Percent of all users	Incorrect	Percent incorrect
AND	832	5	418	50
OR	39	0	11	28
AND NOT	47	0	9	19
( )	120	1	0	0
+ (plus)	826	5	303	30
– (minus)	508	3	362	38
“ ”	1019	6	32	0

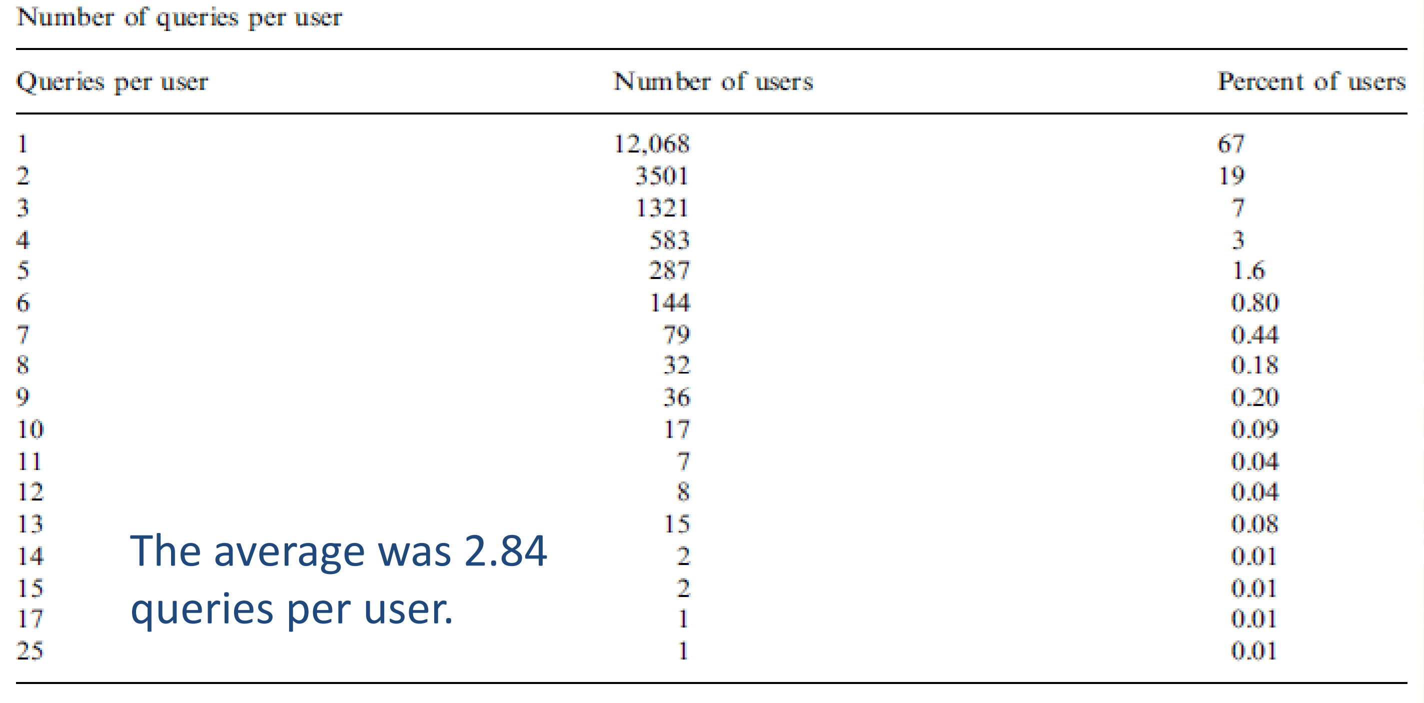
Boolean operators were not used much, with AND receiving the greatest use. These numbers were significantly lower than searches from IR systems.

Number of Queries by User

A number of users went on to either modify their query, view subsequent results, or both. The average session length, ignoring identical queries, was 1.6 queries per user. (majority of users [67%] did not go beyond their first and only query.)

The query length observed is similar to results from other studies. This deviates significantly from traditional IR searching.

The mean number of search terms used in regular IR systems ranged from about 7 to 15. This is about three to seven times higher than this study. (2.21 terms)



Number of queries per user

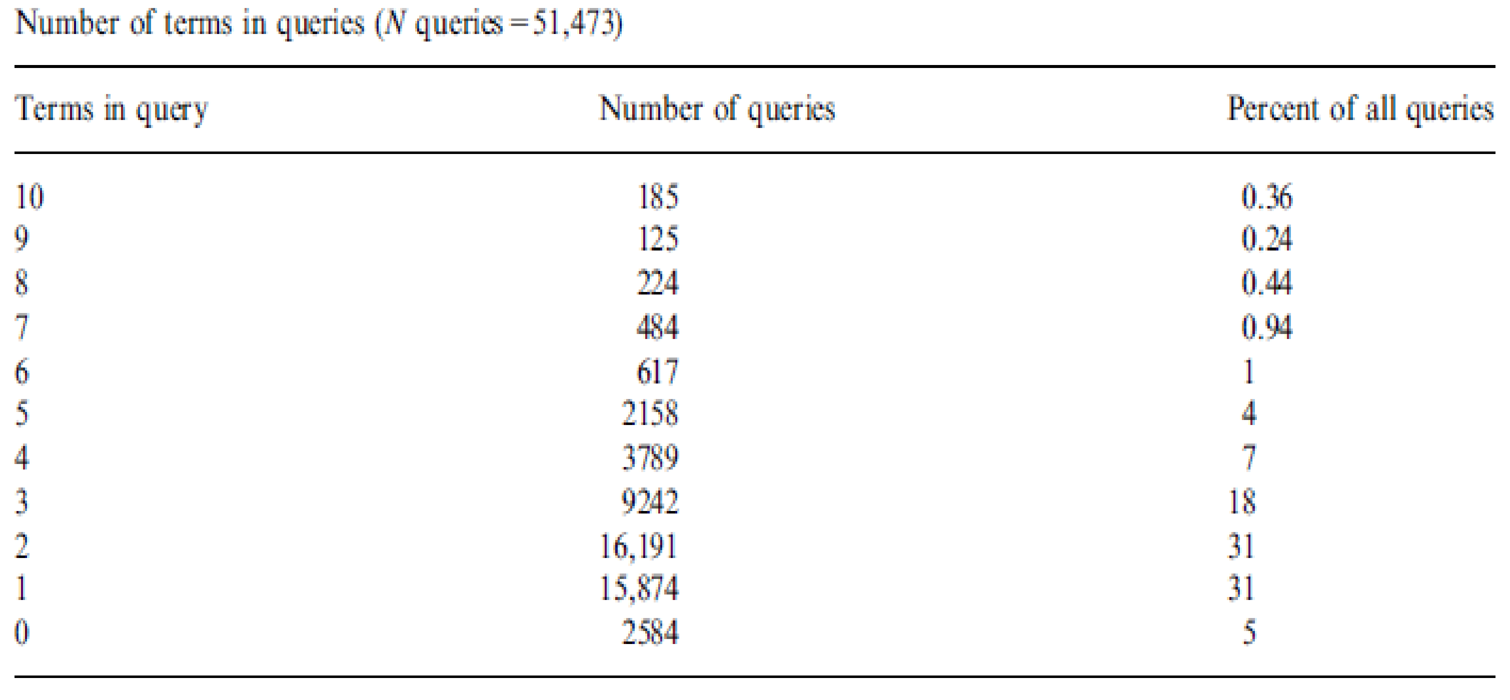
Queries per user	Number of users	Percent of users
1	12,068	67
2	3501	19
3	1321	7
4	583	3
5	287	1.6
6	144	0.80
7	79	0.44
8	32	0.18
9	36	0.20
10	17	0.09
11	7	0.04
12	8	0.04
13	15	0.08
14	2	0.01
15	2	0.01
17	1	0.01
25	1	0.01

The average was 2.84 queries per user.

Number of Terms in Queries

When a user enters a command for relevance feedback (More Like This), the Excite transaction log counts that as a query, but a query with zero terms.

For IR researches, some 11% of search terms came from relevance feedback. Thus, the relevance feedback on the Web is used half as much as in traditional IR searches.

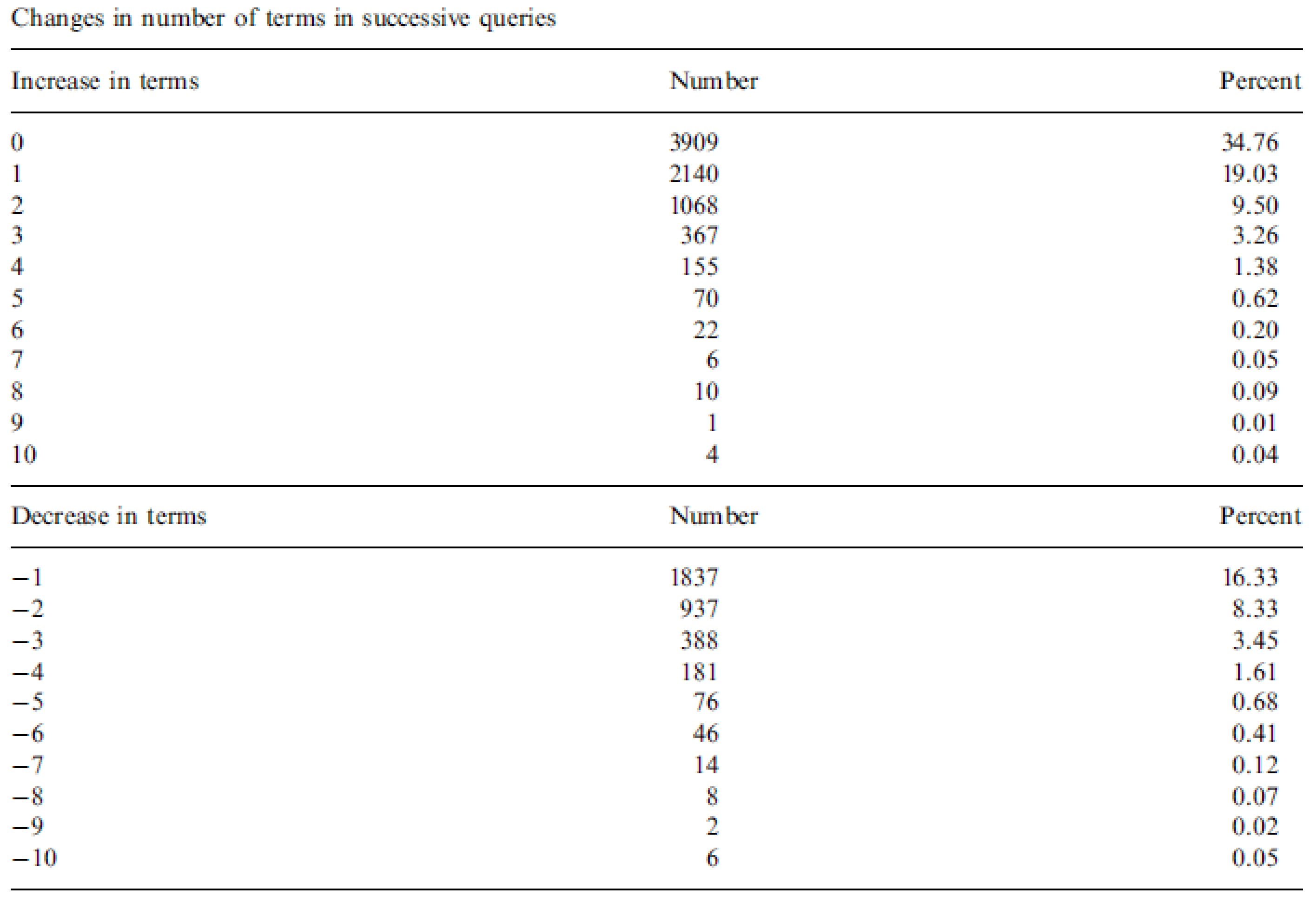


Number of terms in queries (*N* queries = 51,473)

Terms in query	Number of queries	Percent of all queries
10	185	0.36
9	125	0.24
8	224	0.44
7	484	0.94
6	617	1
5	2158	4
4	3789	7
3	9242	18
2	16,191	31
1	15,874	31
0	2584	5

Number of Terms in Queries

How user modified their queries



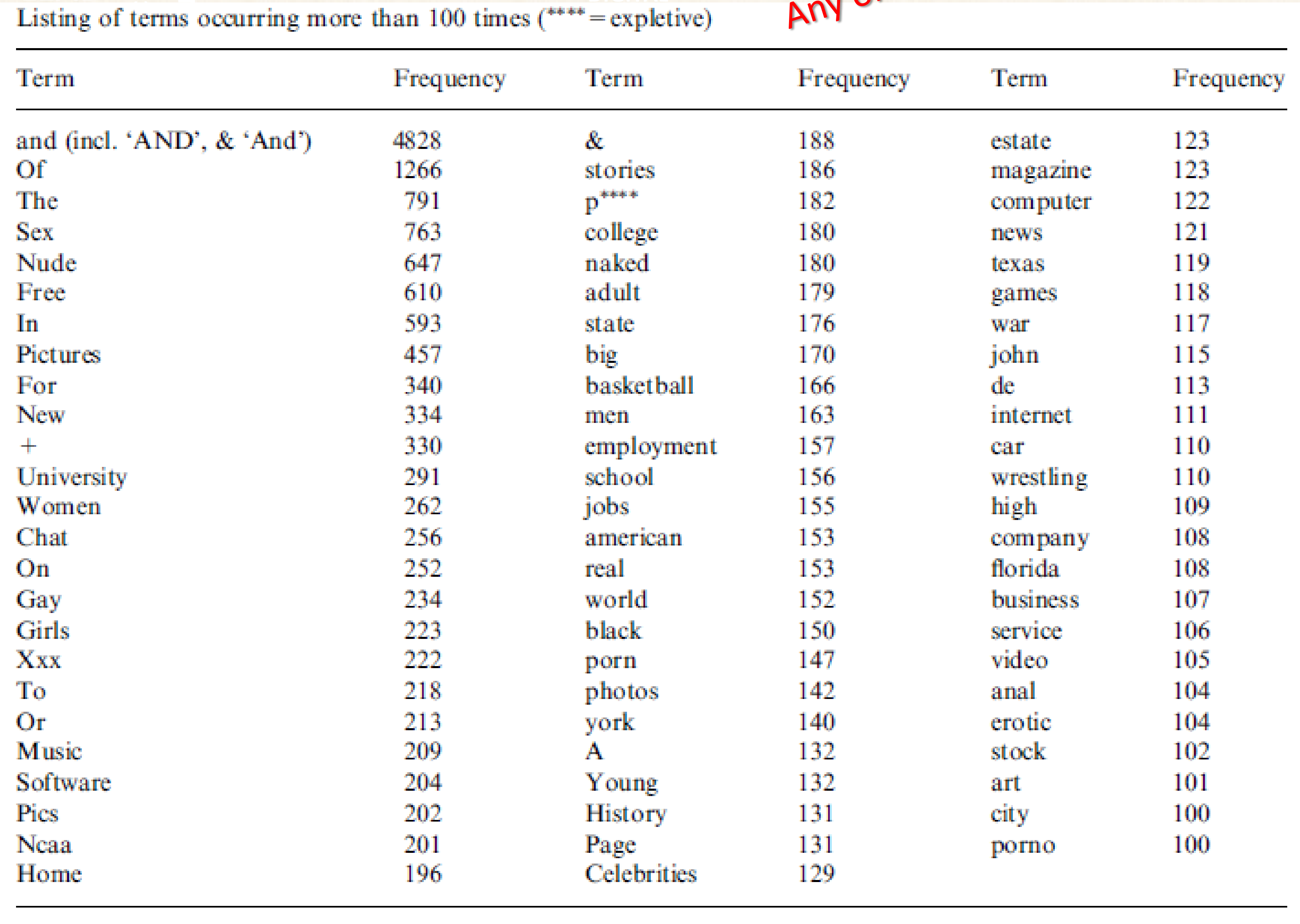
Changes in number of terms in successive queries

Increase in terms	Number	Percent
0	3909	34.76
1	2140	19.03
2	1068	9.50
3	367	3.26
4	155	1.38
5	70	0.62
6	22	0.20
7	6	0.05
8	10	0.09
9	1	0.01
10	4	0.04

Decrease in terms	Number	Percent
−1	1837	16.33
−2	937	8.33
−3	388	3.45
−4	181	1.61
−5	76	0.68
−6	46	0.41
−7	14	0.12
−8	8	0.07
−9	2	0.02
−10	6	0.05

Number of Terms in Queries

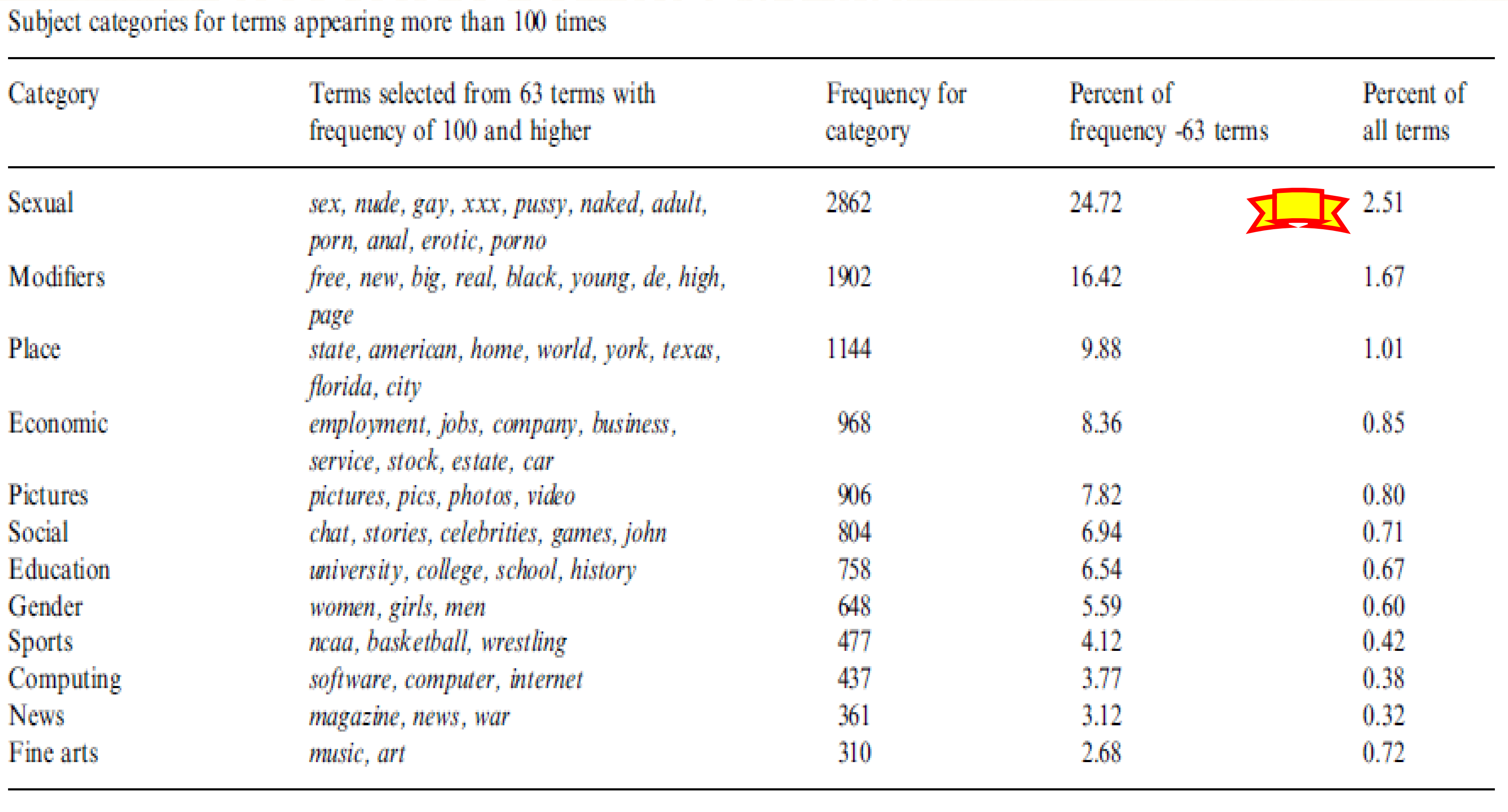
Any one similar to another ?



Listing of terms occurring more than 100 times (**** = expletive)

Term	Frequency	Term	Frequency	Term	Frequency
and (incl. 'AND', & 'And')	4828	&	188	estate	123
Of	1266	stories	186	magazine	123
The	791	p****	182	computer	122
Sex	763	college	180	news	121
Nude	647	naked	180	texas	119
Free	610	adult	179	games	118
In	593	state	176	war	117
Pictures	457	big	170	john	115
For	340	basketball	166	de	113
New	334	men	163	internet	111
+	330	employment	157	car	110
University	291	school	156	wrestling	110
Women	262	jobs	155	high	109
Chat	256	american	153	company	108
On	252	real	153	florida	108
Gay	234	world	152	business	107
Girls	223	black	150	service	106
Xxx	222	porn	147	video	105
To	218	photos	142	anal	104
Or	213	york	140	erotic	104
Music	209	A	132	stock	102
Software	204	Young	132	art	101
Pics	202	History	131	city	100
Ncaa	201	Page	131	porno	100
Home	196	Celebrities	129		

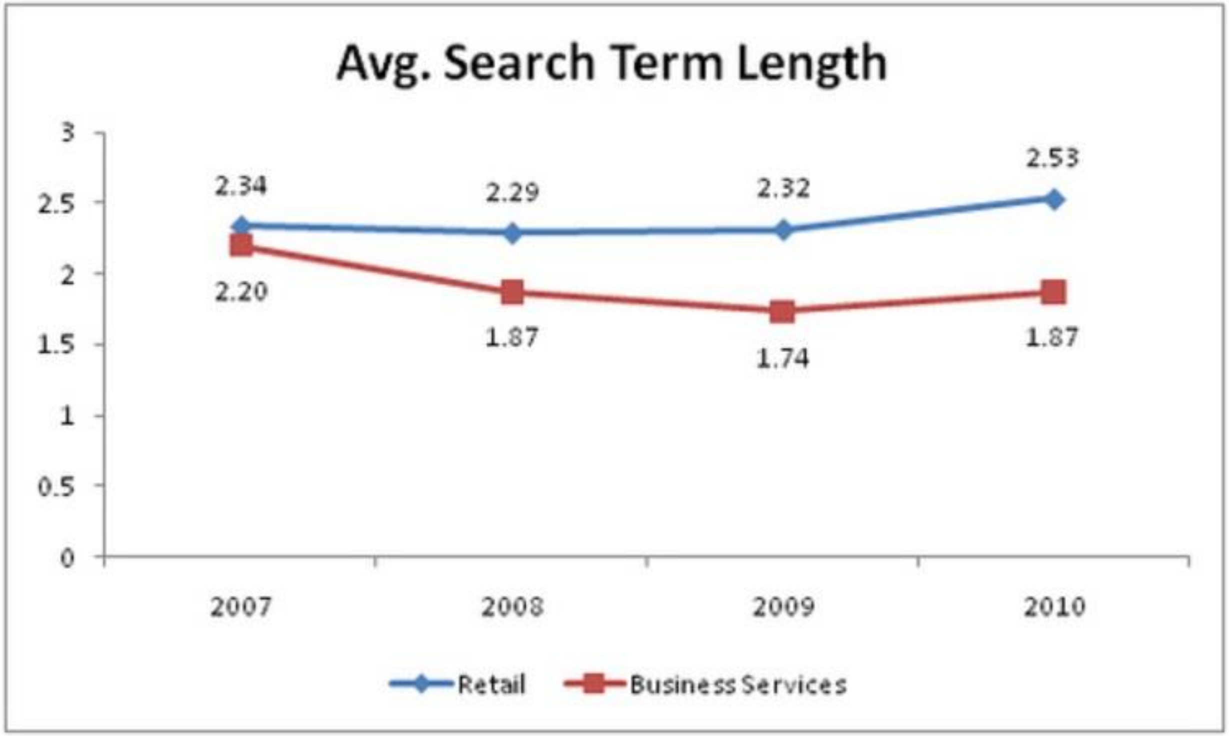
Number of Terms in Queries



Subject categories for terms appearing more than 100 times

Category	Terms selected from 63 terms with frequency of 100 and higher	Frequency for category	Percent of frequency -63 terms	Percent of all terms
Sexual	*sex, nude, gay, xxx, pussy, naked, adult, porn, anal, erotic, porno*	2862	24.72	2.51
Modifiers	*free, new, big, real, black, young, de, high, page*	1902	16.42	1.67
Place	*state, american, home, world, york, texas, florida, city*	1144	9.88	1.01
Economic	*employment, jobs, company, business, service, stock, estate, car*	968	8.36	0.85
Pictures	*pictures, pics, photos, video*	906	7.82	0.80
Social	*chat, stories, celebrities, games, john*	804	6.94	0.71
Education	*university, college, school, history*	758	6.54	0.67
Gender	*women, girls, men*	648	5.59	0.60
Sports	*ncaa, basketball, wrestling*	477	4.12	0.42
Computing	*software, computer, internet*	437	3.77	0.38
News	*magazine, news, war*	361	3.12	0.32
Fine arts	*music, art*	310	2.68	0.72

Trends In Search Query Length



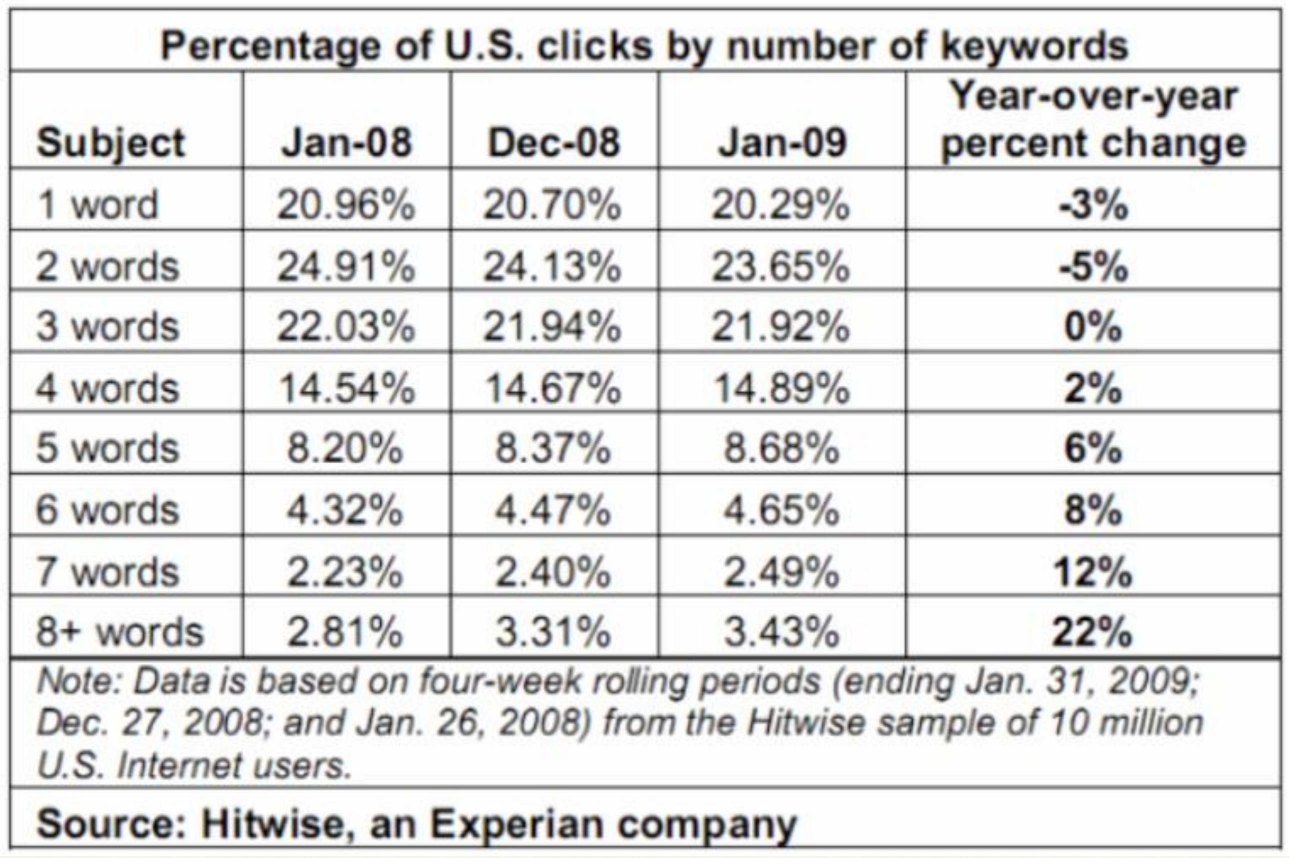
http://searchengineland.com/caution-reported-trends-in-search-query-length-may-be-misleading-41641

Reference by the book **Transforming Technologies to Manage Our Information: The Future of Personal Information Management**

Trends In Search Query Length

If queries are getting longer, the long tail of search must be increasing?

If web users are getting more specific in their searches, marketers need to keep adding increasingly specific search terms in campaigns?



Percentage of U.S. clicks by number of keywords				
Subject	Jan-08	Dec-08	Jan-09	Year-over-year percent change
1 word	20.96%	20.70%	20.29%	-3%
2 words	24.91%	24.13%	23.65%	-5%
3 words	22.03%	21.94%	21.92%	0%
4 words	14.54%	14.67%	14.89%	2%
5 words	8.20%	8.37%	8.68%	6%
6 words	4.32%	4.47%	4.65%	8%
7 words	2.23%	2.40%	2.49%	12%
8+ words	2.81%	3.31%	3.43%	22%

Note: Data is based on four-week rolling periods (ending Jan. 31, 2009; Dec. 27, 2008; and Jan. 26, 2008) from the Hitwise sample of 10 million U.S. Internet users.

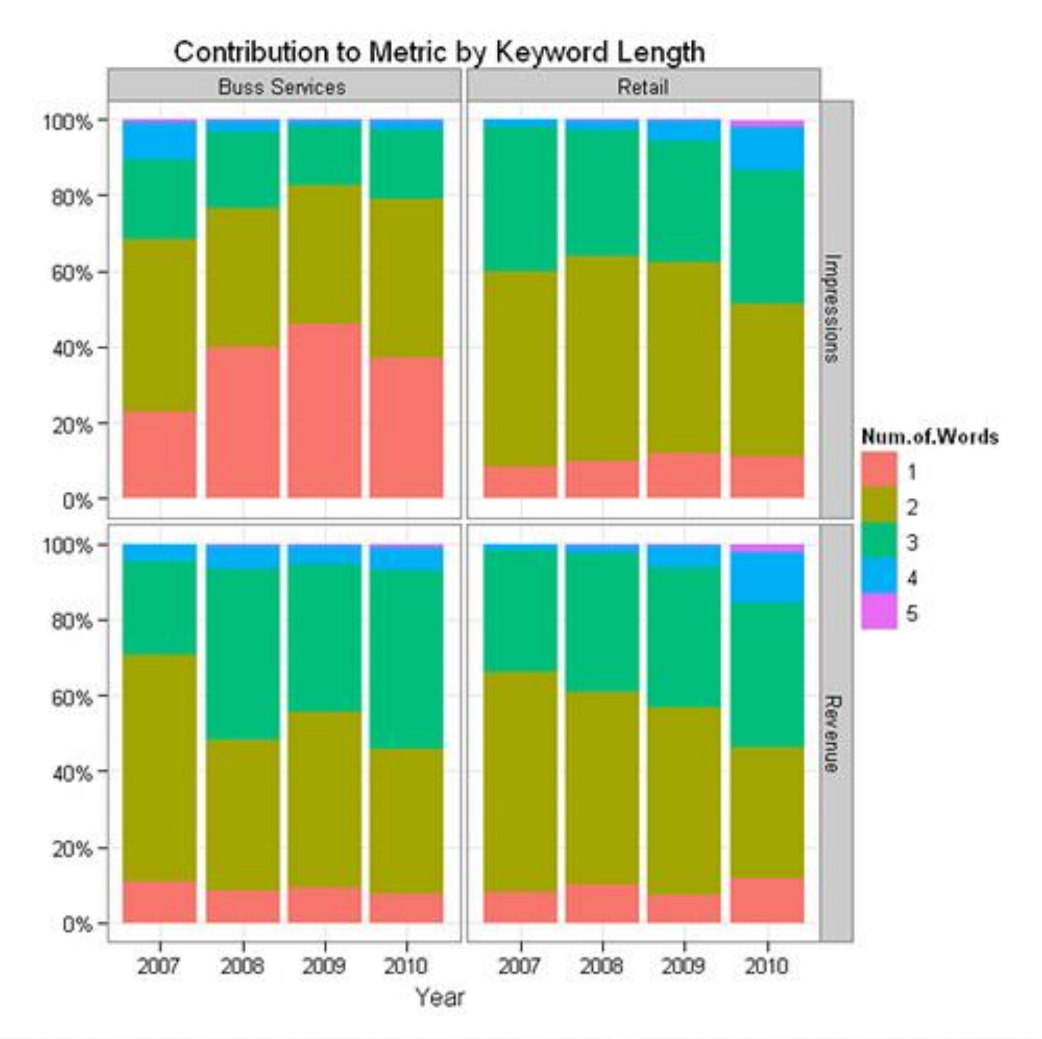
Source: Hitwise, an Experian company

Trends In Search Query Length

Search terms of length greater than 5 words contributed less than 1% of volume both in impressions and revenue.

4 and 5 word search queries was less than 3% of volume in 2007 whereas they now account for more than 15% of volume by retail.

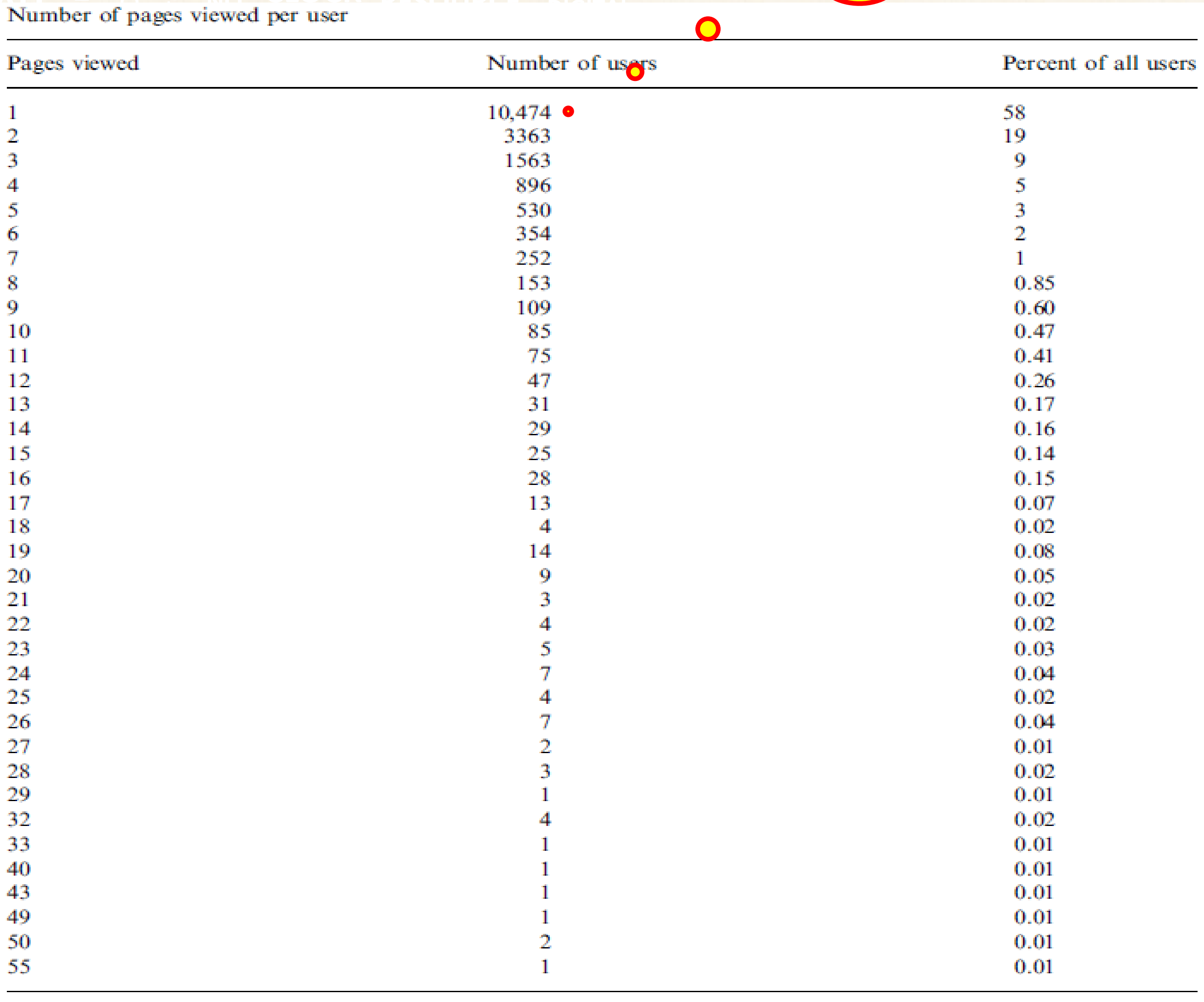
However, in the business services vertical, there was no such trend, and one and two word search terms have been increasing.



http://searchengineland.com/caution-reported-trends-in-search-query-length-may-be-misleading-41641

Viewing of Results

Done?



Number of pages viewed per user

Pages viewed	Number of users	Percent of all users
1	10,474	58
2	3363	19
3	1563	9
4	896	5
5	530	3
6	354	2
7	252	1
8	153	0.85
9	109	0.60
10	85	0.47
11	75	0.41
12	47	0.26
13	31	0.17
14	29	0.16
15	25	0.14
16	28	0.15
17	13	0.07
18	4	0.02
19	14	0.08
20	9	0.05
21	3	0.02
22	4	0.02
23	5	0.03
24	7	0.04
25	4	0.02
26	7	0.04
27	2	0.01
28	3	0.02
29	1	0.01
32	4	0.02
33	1	0.01
40	1	0.01
43	1	0.01
49	1	0.01
50	2	0.01
55	1	0.01

Viewing of Results

The mean number of pages examined per user was 2.35. Most users, 58% of them, did not access any results past the first page.

Were they so satisfied with the results that they did not need to view more?

Were a few answers good enough?

Is the precision that high?

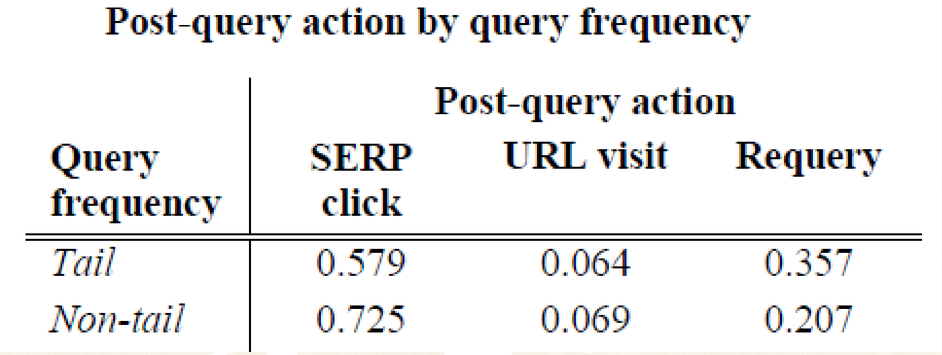
Are the users after precision? Or

Did they just give up and get tired of viewing results?

User Behavior for Rare Versus Common Queries & Goals

Differences in the behavior of searchers with changes in the rarity of queries

Following tail queries, SERP clicks are less common (0.579 vs. 0.725) and requeries are more common (0.357 vs. 0.207), both of which indicate that the results returned by the search engine were not as useful to searchers.



Post-query action by query frequency

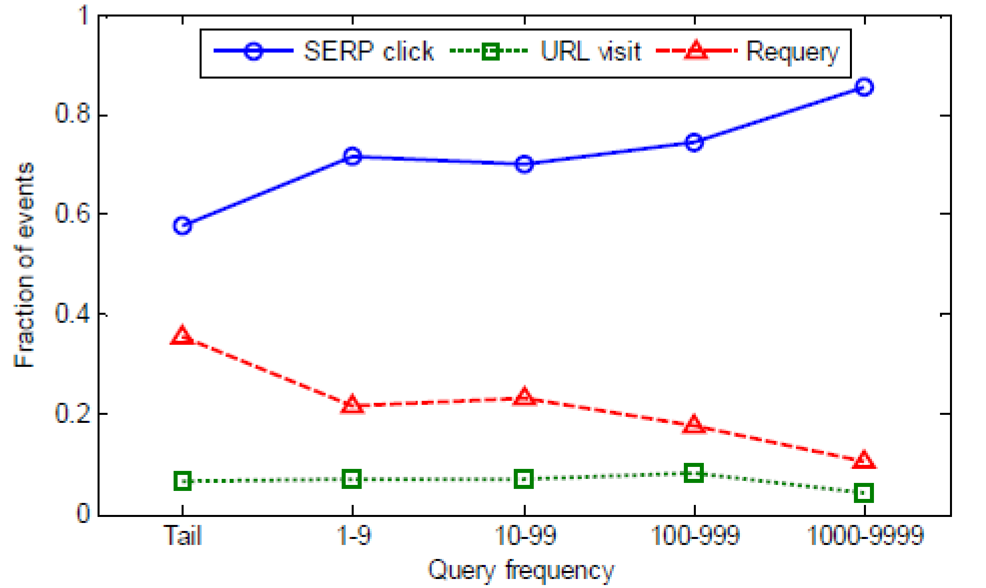
Query frequency	Post-query action SERP click	 URL visit	 Requery
Tail	0.579	0.064	0.357
Non-tail	0.725	0.069	0.207

User Behavior for Rare Versus Common Queries & Goals

Differences in the behavior of searchers with changes in the rarity of queries

The figure below shows that SERP clicks increase and requeries decrease smoothly as the frequency of queries increases.

Besides, search engines are not doing as good a job of satisfying searchers for rarer queries as they do on more common ones



Post-query action by query frequency bins

Summary

Internet IR vs. Traditional IR

Different search term mean number

Regular IR system range from about 7 to 15 search terms but internet IR is 2.21 terms.

Relevance feedback was rarely used.

About one in 20 queries used the feature “More Like This”. In comparison with professionally assisted IR searching, relevance feedback is apparently used only half as much on the Web.

Boolean operators were seldom used.

One in 18 users used any Boolean capabilities and when using them they have great difficulty in getting them right.

Summary

Trends In Search Query Length

Query length getting longer but not too many difference

The average search term length increased for the retail but decreased in the business services vertical.

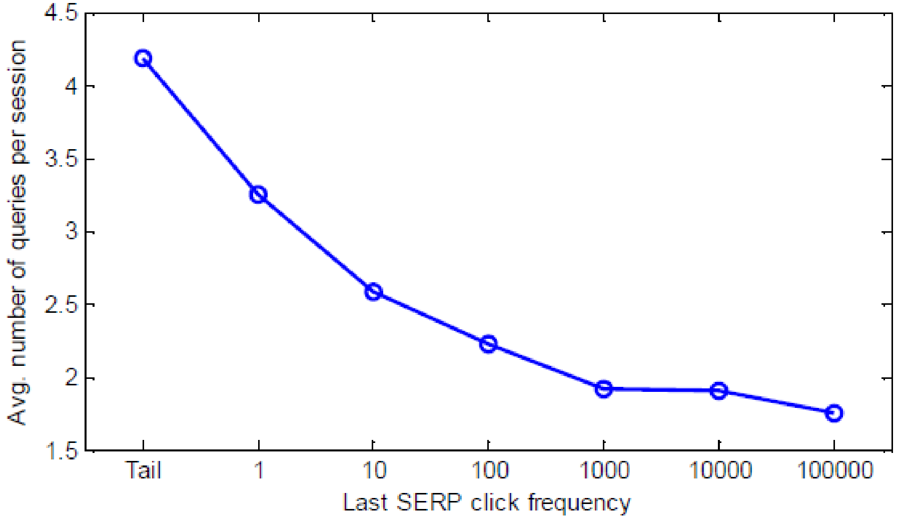
As the business services example, the results may be contrary to what we expect.

- Average search term for retail (B to C) is 2.53 in 2010.
- Average search term for business service (B to B) is 1.87 in 2010.

Summary

User Behavior for Rare Versus Common Queries

SERP clicks increase and requeries decrease smoothly as the frequency of queries increases.



Average number of queries by frequency of information goal

Thank you