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.

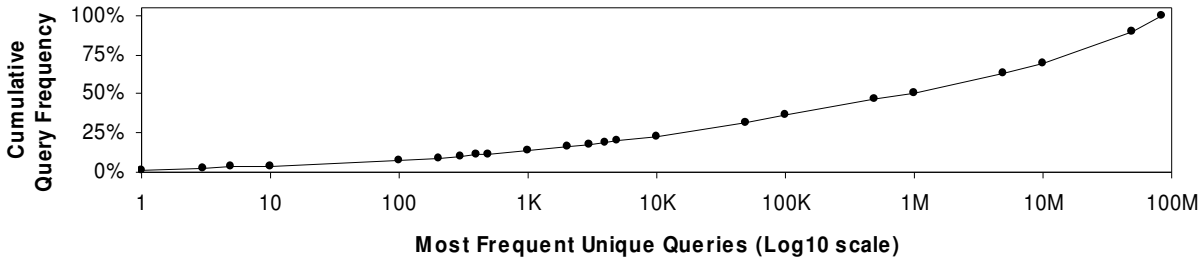
In each section, the graphical measures themselves comprise the majority of the sectional content[1], with supplementary text given in the form of either graphic annotations or short summaries of the section as a whole. We have chosen this style of presentation for several reasons. Foremost, effective 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.

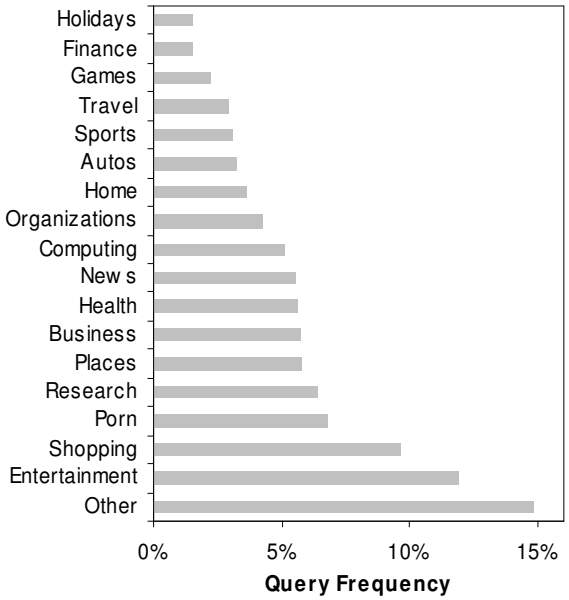
[1] To assure legibility, this paper requires a 600 dpi (or greater) printer.

2. QUERY SPACE

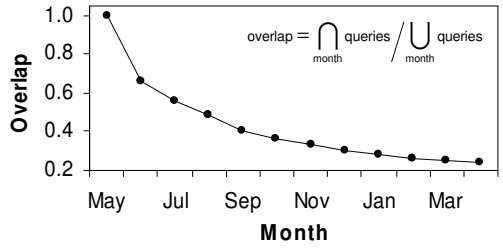
(2.1)



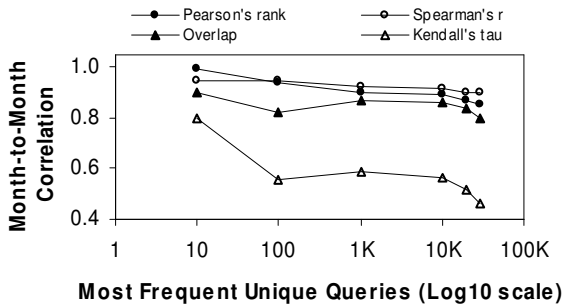
(2.2)



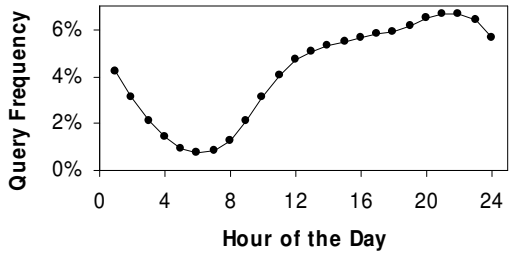
(2.3)



(2.4)



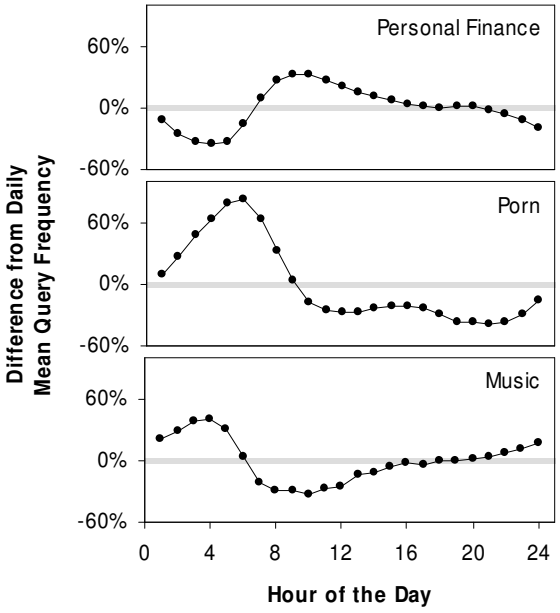
(2.5)



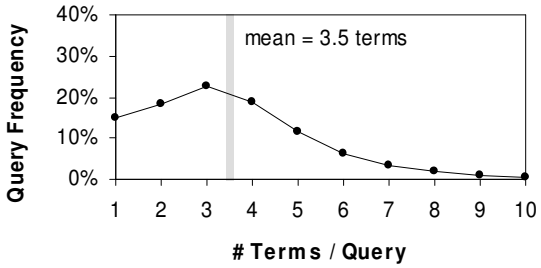
(2.6)

(2.7)

(2.8)



(2.9)



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), expressed in millions of variations (2.1).

3. USER SESSIONS

(3.1)

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

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

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

(3.2)

In this session, the user formulates a series of queries in pursuit of multiple tasks.

In general, the average series of query formulations within a user session can be summarized as a probability matrix (3.4) between the following formulation states:

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

Timeline (hh:mm:ss)		Query
00:00	◯	dail news
01:06	(c)	daily news
10:42	◯	frito lay
13:48	◯	smoking celebrities
14:36	(>)	smoking celebrities
22:18	◯	cd reviews
32:48	(>)	cd reviews
40:06	◯	bestbuy.com
41:18	◯	tower records
47:00	◯	money making opportunities
51:42	◯	gumball machines
51:54	(>)	gumball machines
57:54	(>)	gumball machines
01:03:48	◯	vending opportunities
01:05:48	◯	inventions
01:09:00	(>)	inventions
01:18:36	◯	patents
01:23:12	(<)	smoking celebrities
01:33:18	◯	images.mp3.com
01:33:36	◯	www.ajolie.com
01:36:24	◯	the sopranos
01:38:30	(>)	the sopranos

(3.3)

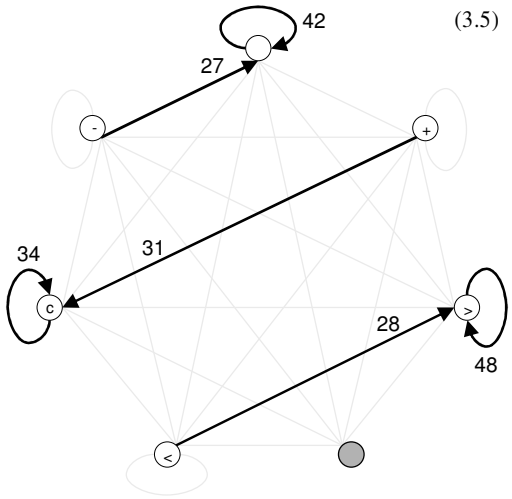
Navigational queries account for 21% of the total query frequency.

Timeline (days)		Query
0	◯	google
2	◯	yahoomail
8	◯	travellodge
13	(<)	yahoomail
24	◯	www.trapeze.com

(3.4)

Probability from State \ To State	%	◯	(+)	(-)	(c)	(>)	(<)	●
◯		**42**	6	2	15	24	6	5
(+)		25	4	3	**31**	26	8	4
(-)		**27**	18	2	15	26	8	4
(c)		20	4	3	**34**	28	6	5
(>)		20	5	1	17	**48**	5	4
(<)		27	4	1	13	**28**	21	6

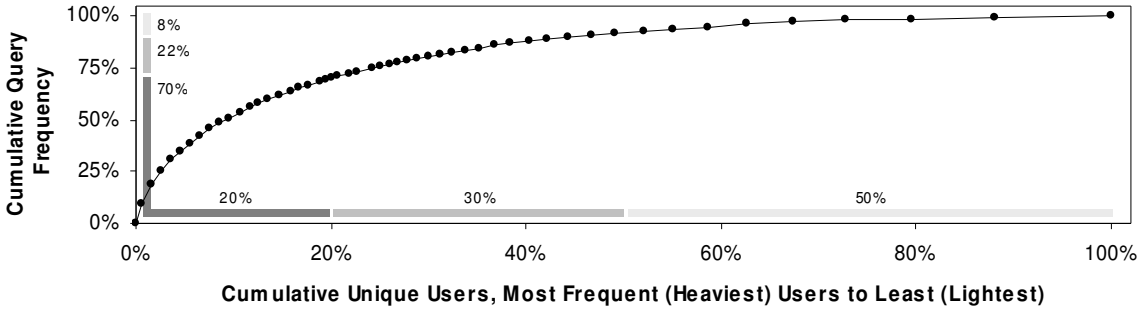
(3.5)



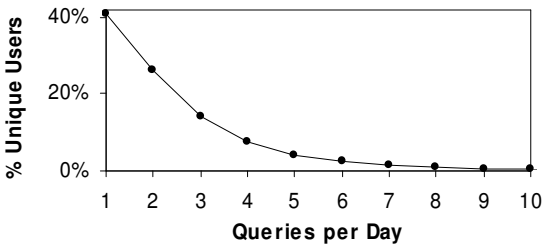
On a given day, 41% of users search just once. Such user behavior is described in the following section.

4. USER BEHAVIOR

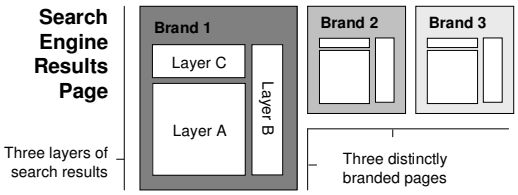
(4.1)



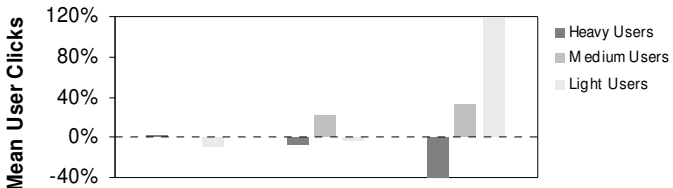
(4.2)



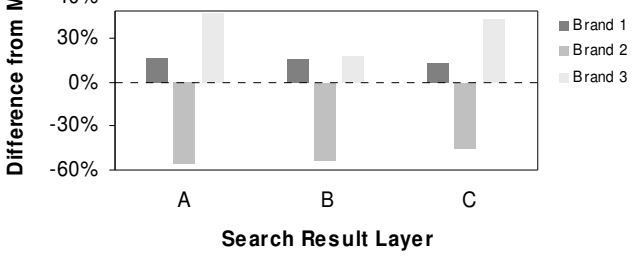
(4.3)



(4.4)



(4.5)



(4.6)

Search Engine	Precision @ 10	Mean Avg. Precision @ 10
Meta-engine 1	.742	.693
Meta-engine 2	.704	.655
Engine 1	.685	.625
Engine 2	.668	.607
Engine 3	.666	.599
Engine 4	.666	.598
Engine 5	.661	.592
Engine 6	.652	.577
Engine 7	.625	.571
Engine 8	.628	.567
Engine 9	.629	.555
Engine 10	.613	.537

(4.7)

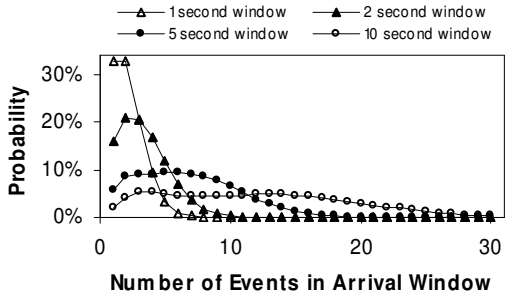
Page Rank	% Total Views	Result Rank	% Total Clicks
1	86	1	45
2	6	2	13
3	2	3	9
4	1	4	6
5	1	5	5
6	1	6	4
7	1	7	3
8	< 1	8	3
9	< 1	9	2
10	< 1	10	3
> 10	< 1	> 10	9

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.

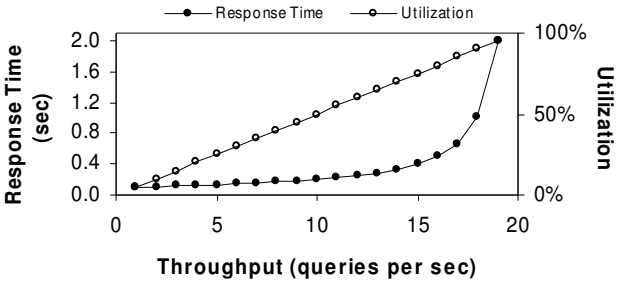
5. OPERATIONAL REQUIREMENTS

(5.1)

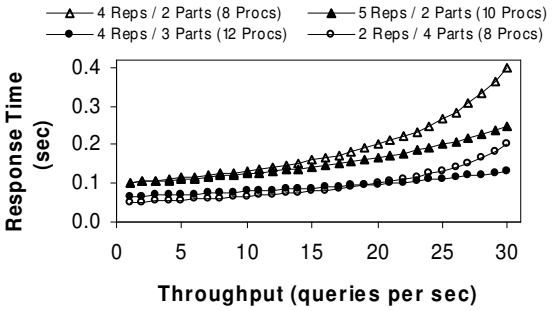


Given a query arrival process (5.1), a document collection, and the relationship between collection size and service time ⟋, queuing network theory can be used to derive the response time and utilization at varying throughputs (5.2) for a given system architecture (5.4) consisting of replications and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

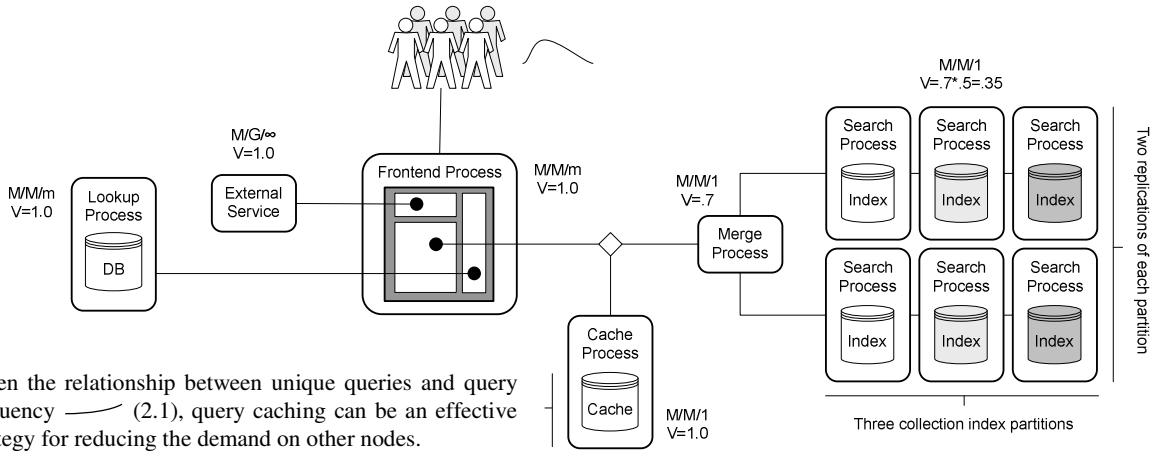
(5.2)



(5.3)

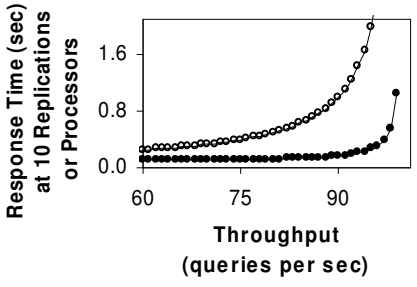


(5.4)

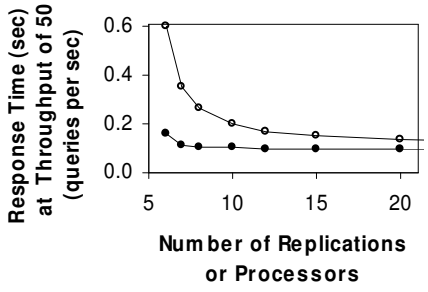


Given the relationship between unique queries and query frequency ⟋ (2.1), query caching can be an effective strategy for reducing the demand on other nodes.

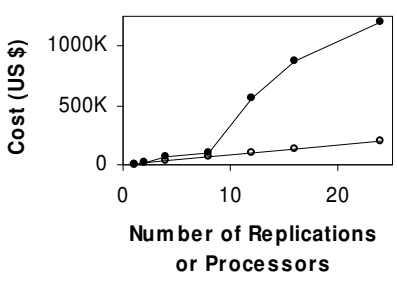
(5.5)



(5.6)



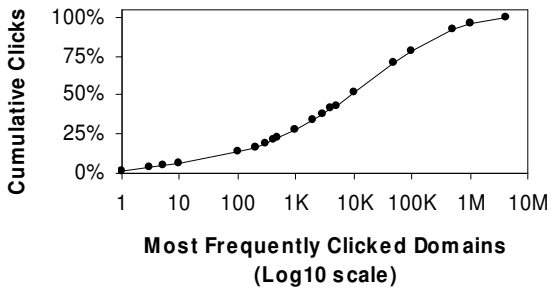
(5.7)



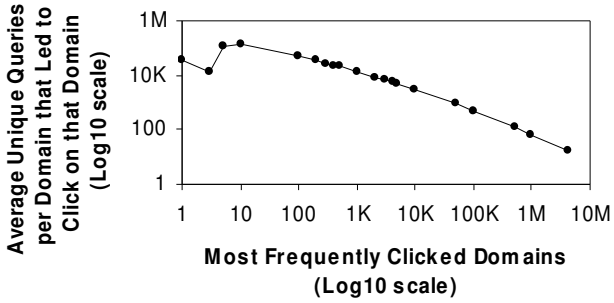
Multiprocessor systems can optimize queue distributions for significant performance gains (5.5, 5.6). Trends in hardware costs, though, suggest a different solution (5.7).

6. CONTENT SPACE

(6.1)



(6.2)



(6.3)

Intersection of Top 10 Search Results for Search Engines E1 – E10

Mean intersection = 37%

%	E1								
E2	33	**E2**							
E3	31	89	**E3**						
E4	31	78	82	**E4**					
E5	34	80	81	79	**E5**				
E6	27	25	23	24	25	**E6**			
E7	30	77	78	79	76	24	**E7**		
E8	29	25	23	24	25	24	24	**E8**	
E9	26	22	20	21	22	25	20	22	**E9**
E10	26	26	24	24	25	21	24	21	18

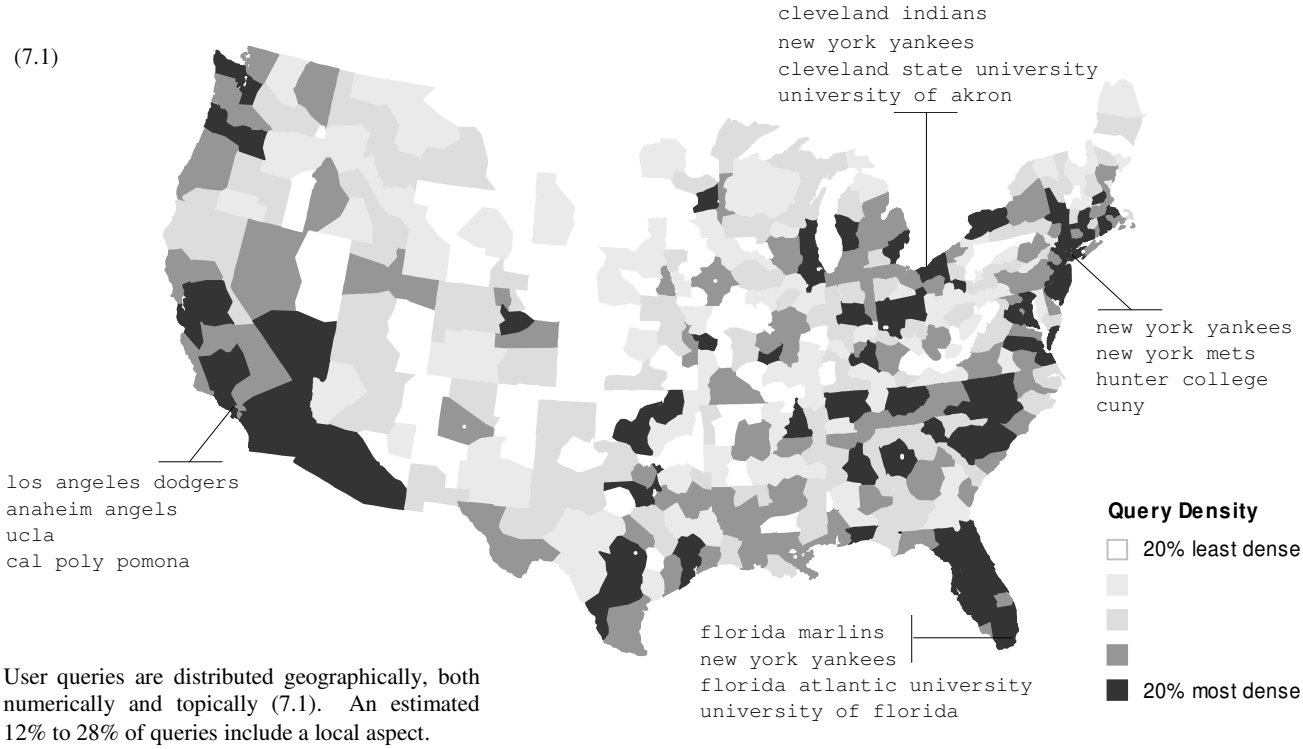
Of an estimated 50 million web domains, less than 1% account for half of all user clicks via search results (6.1).

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

(7.1)



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, Shawn 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," *JASIST*, 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 Metasearch 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.