Searching the Web and more — a juxtaposition of online search traces*

Brian D. Davison and Wei Zhang
Department of Computer Science and Engineering, Lehigh University
19 Memorial Drive West, Bethlehem, PA 18015
{davison,wez5}@cse.lehigh.edu

Abstract

The information retrieval task is larger than the problem of searching for documents on the Web. In this paper we broaden our analysis to include search logs of many Web search engines, peer-to-peer query logs, newsgroup search logs, and queries to FTP archives. We calculate and compare the characteristics of each of these query logs from 2003 to find commonalities and differences across a wide spectrum of online query workloads. We found Boolean operator usage to be rare; much longer queries in peer-to-peer traffic than Web; that searchers click on slightly more than two results per query; that peer-to-peer and FTP logs are more likely to include file-type extensions; and, that caching of query results is likely to be of value for both WWW and peer-to-peer traffic.

1 Introduction

In recent years, much information retrieval research has focused on the problems raised by searching for documents on the Web. Indeed, the unprecedented growth of Web content, and the wide demand for effective search has made it an important problem to consider. However, the information retrieval task is much larger than the problem of Web search. In peer-to-peer file-sharing networks, which have displaced Web traffic as the leading source of network traffic [14, 27], a primary concern is how to find desired content in networks offering terabytes of shared data.

*Technical Report LU-CSE-05-005, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015.

In this paper we broaden our view (slightly) from strictly Web search to include analyses of query logs from peer-to-peer systems, and alternative online content, such as newsgroups, FTP archives, and MSN's automatic search feature for Internet Explorer. Our goal in this work is to look for commonalities across a broader selection of query logs, and to find differences between sources of logs to better equip IR researchers with intuitions about additional query workloads. A better understanding of query stream characteristics is important for search engine design, both in terms of quality (perhaps providing better results for a particular domain at the expense of other, less-popular domains of interest) and performance in terms of response time or throughput. Knowing, for example, that long queries are the norm might have a significant influence on the chosen implementation.

Unlike in-depth studies of a single large trace [23, 21], we are able to compare results across traces, by using a consistent infrastructure to measure query log characteristics. Past Web studies have found that users generate short queries, in contrast to traditional information retrieval searching studies in which the number of search terms range from 7-15 [8]. Our broader traces span a wider range, with median query lengths of 1-6 terms. Most of the traces we examine are recent, from 2003.

In the remainder of this paper, we describe our data sets and how they were collected. We then review related work. We analyze various characteristics of the data sets, including query lengths, term popularities, temporal locality, file types found, and click-through analysis. We conclude with a summary of our contributions.

2 Data Sets

Our logs are from a variety of sources. We have three principal sources, which we describe below. Regardless of the source, some amount of cleansing and normalization was required. Across all traces we normalized the case of all queries, and reduced consecutive white space to a single space.

2.1 AltaVista 2001 sample

The first data source is a set of over 7 million queries submitted to AltaVista starting on September 28, and ending October 3, 2001. This trace has been analyzed elsewhere [15]. Unfortunately, this log does not contain the entire stream of queries that was submitted on those dates. Instead, it contains a sample of the search phrases (but includes all instances of the selected phrases). As a result, examination of this log does not directly answer

questions such as the most frequent query sent to AltaVista. As a result of the sampling and a lack of user identification, this log does not permit session-based analysis.

We include this large but older query log primarily as a reference against which the newer data can be compared.

2.2 Proxy cache queries

The second trace contains queries sent through NLANR's IRCACHE Web proxy cache infrastructure [19, 28], starting in late January 2003 through December 2003. This infrastructure service is provided at no charge to proxy cache operators throughout the world. As an example, on December 1, 2003, the IRCACHE proxy infrastructure served more than 4.76 million requests, corresponding to 38GB of content. A quarter of that content was served from cache, rather than being fetched from the origin server. (See [5] for a tutorial on caching for the Web.)

For our purposes, the logs from this infrastructure provide insight into the searching activities of users across a variety of search systems. Queries in this log include those sent to Google, AltaVista, MSN, HotBot, AllTheWeb, and Yahoo. Queries are also captured to non-traditional search services, like Google Groups which is a searchable USENIX Newsgroups archive (34 thousand queries), AllTheWeb's FTP search service (342 thousand queries), and auto.search.msn.com (termed Auto-MSN) which receives "queries" that are typed into the address bar of Internet Explorer 6 (178 thousand queries), and includes incorrect hostnames that are typed there (e.g., www.yahoo.co).[1]

These logs were generated by matching known search engine hostnames and quer yformats to URLs used in requests. As a result, there are many URL requests present in the logs that do not correspond to an actual query. In some cases, these requests are still useful, as in the logging of click-throughs. In others, the request generates a page but the content is not specific to a textual query. Recognizable but ignored requests included viewing newsgroup postings, some image URLs, and browsing a topic hierarchy. Out of the original log of 3.63 million requests, our code recognized 3.43 million queries, and 287k click-throughs. However, a number of those queries were to country-specific engines (e.g., `uk.altavista.com`), or to other sites but with too few queries, so they were excluded from our analysis.

Out of the 3.26 million standard Web search queries extracted from the IRCACHE logs, the bulk (76.6%) were to Google. Yahoo was next, with

[1] This behavior is built-in to Internet Explorer 6; IE5 had options to turn it on or off, and IE4 sent the queries to randomly selected search engines.

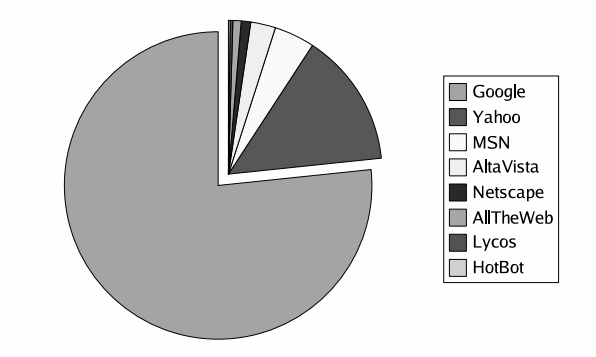


Figure 1: Relative presence of the major engines in the IRCACHE dataset.

14.3%, followed by MSN with 4.3%. The full distribution is shown in Figure 1. While others have published relative market shares and searches per day [26, 25], they are rarely comparable, as a result of being based on different definitions of what sites are included, which user market is counted, etc.

2.3 Peer-to-peer queries

We have collected queries from two peer-to-peer file sharing applications. The first is Gnutella [12], a strictly peer-to-peer file sharing application. A client application can share files specified by the user, and match those files to queries that are received from other users on the network. According to statistics published by LimeWire.com (a publisher of one popular Gnutella client), the Gnutella network had more than 200,000 users online as of January 2004. We have used a peer-to-peer gateway between the Web and Gnutella, enabling Gnutella users to issue Web search queries and retrieve content from the Web [6]. We logged 1,598,070 Gnutella queries from sources within six network hops that were received by this gateway between November 3, 2003 to January 19, 2004.

The second P2P application logged was eDonkey. eDonkey is a file-sharing P2P network, created initially through the development of the eDonkey2000 client by MetaMachine in 2000. Like Napster and in contrast to Gnutella, eDonkey requires servers to index the files that are shared.[2] As of early January 2004, slyck.com reported 1.56 million users of eDonkey (vs. 184k for Gnutella and 3.4M for FastTrack). We operated a small eDon-

[2]Such a design introduces some scalability issues and as a result, in 2002, MetaMachine introduced OverNet, which uses many of the same protocols, but no centralized servers.

key server for up to 1000 simultaneous users and collected 142,013 queries sent to it between December 28, 2003 and January 17, 2004.

3 Related Work

In this section, we consider prior work, reviewing those who have analyzed query characteristics, search engine click-through logs, and temporal locality in query logs that can be exploited through caching.

3.1 Analysis of query characteristics

Clip2 DSS [3] analyzed early Gnutella traffic. They manually cataloged 2000 queries recorded on September 19, 2000, and found that more than 32% were for music artists, more than 26% gave a file extension only, about 9% were for a song title or song title and artist, almost 9% were automated crawlers/indexers, about 7% were for adult themes, over 3% were for software, and more than 2% were for movies. Examining the file extensions more closely, they tested a 150,000 query log and found close to 42% were for video files, more than 39% were for audio files, 4% for image files, and 12% for software.

Silverstein et al. [21] analyzed a very large 1998 AltaVista query log, consisting of approximately a billion queries captured over 43 days.

Jansen et al. [9, 10] analyzes a small (51k query) 1998 Excite trace. They investigated query length, query structure (use of operators), term occurrence and the user session. They found that 10.3% of all queries used a Boolean operator. Phrases were used in about 5% of all queries.

Spink et al. [23] studied a 1997 log of over one million Excite web queries by more than 200k users. They investigated query length, term co-occurrence, query structure and session activities. Their work shows that most Web queries are short (mean 2.16, median 2), infrequently modified and simple in structure (5% used phrases, and 7% used + or - operators).

Spink et al. [22] compares the searching behaviors of users of two major search engines in the US and Europe respectively. They performed the comparison in terms of query length (mean of 2.6 terms), topic category and top terms. We examine similar characteristics of the datasets described in this paper.

Ling et al. [16] studied an unspecified commercial knowledge-based search engine to find generalized query patterns, telling editors what topics users are interested in. They analyzed two days' worth – 260k queries (along with click-throughs) and 20,000 articles. They found that most searchers

used a small number of keywords — 52.5% of queries used only a single term; 32.5% used two terms, 10% used three, and 5% used four or more.

Cacheda and Vina [2] report statistics on the first screen of results only, from a two-week log of a Spanish portal in May 2000. They found that 14.63% of the queries used advanced operators and 3.64% included phrases.

See Jansen et al. [7] for a review of additional studies.

3.2 Click-through data

Many commercial search engines collect click-through data. Google samples click-through data for internal evaluation. One that used it directly to present results based on user activity is Direct Hit (purchased by Ask Jeeves in 2000, and whose technology is now incorporated into Teoma [Apostolos Gerasoulis, personal communication, 2003]).

Ling et al. [16] report that most users clicked on one article per query — only .06% clicked on more than one.

Click-through information can be used to improve or augment search engine performance. Beeferman and Berger [1] use click-through data to find related queries. Cui et al. [4] examine Encarta query logs for query expansion. Joachims [11] examined query logs that also contained click-through data. Joachims demonstrated the ability to learn better ranking functions, using a specially collected log within a meta search engine environment.

3.3 Caching analysis

There is a tangible cost (both in time and money) in calculating the results of a query. However, those costs can be amortized across as many query instances that arrive while the results are stored (cached) and are still valid. Here we review prior studies of query traffic and the inherent locality found.

3.3.1 WWW traffic

A number of studies have examined the temporal locality and cachability properties of search engine query logs [21, 17, 20, 29, 15]. Silverstein et al. [21] found that more than 60% of the queries seen were never repeated. Popular queries were repeated many times — the top 25 most popular queries represented 1.5% of all traffic. Xie and O'Hallaron [29] examined the locality of queries in a 2001 Vivisimo log and a 1999 Excite log to determine where caching should occur. They found that queries follow a zipf distribution, and that caching should be performed both at the user and at the search engine side to take advantage of the temporal localities of

different workloads. Markatos [17] studied an Excite 1997 query log, finding, among other things, that successive submissions of the same query were found in close proximity. Lempel and Moran [15] examined the 2001 AltaVista trace examined in this work to demonstrate the performance of a probabilistic scheme for caching and prefetching search engine results. They showed that prefetching can improve hit ratios by 50% or more.

3.3.2 P2P traffic

A few researchers have examined the characteristics of peer-to-peer query traffic, primarily focusing on Gnutella. Krishnamurthy et al. [13] plotted query popularities from a Gnutella trace, visually showing a non-Zipfian curve. Sripanidkulchai [24] captured approximately five days worth of Gnutella query traffic (recording more than eight million query instances) and showed that the distribution of the popularity of Gnutella queries has two distinct phases — very popular documents are roughly equally popular, but less popular documents have a distribution which follows a Zipf-like distribution. Markatos [18] captured an hour of traffic from three locations in October 2001. He found that caching of query results for 30 minutes can reduce query traffic by approximately 50%. Zeinalipour et al. [30] captured Gnutella traffic in the summer of 2002. They were able to classify users into three categories (seasonal-content searchers, adult-content searchers, and file extension searchers).

4 Analysis Results

Here we report the results of analyzing the various query logs. We detail statistics of and peculiarities of query lengths, query and term popularities, advanced query usage, temporal locality, file-type distributions, and click-through rates.

4.1 Query lengths

Query representation varies, depending (at least) on the target of the query. In Table 1, we show that the standard Web engines have mean query lengths anywhere from 1.8 to 3.1, and generally a median length of 2 (except for MSN, which is shorter at 1). The Auto-MSN trace is shorter, with a mean length of just 1.3. The non-Web traces extend the opposite range – 50% of Gnutella queries are 4 terms or less, with an average of 5.2, and eDonkey queries are even longer, at 6.4 terms on average, and a median of 6 terms.

Engine	median terms	mean terms	mean bytes
AllTheWeb-FTP	1	1.045	15.01
Auto-MSN	1	1.273	16.44
MSN	1	1.809	16.80
AllTheWeb	2	1.967	15.62
Lycos	2	2.128	15.58
Netscape	2	2.201	18.94
HotBot	2	2.440	17.63
Yahoo	2	2.417	17.36
GoogleGroups	2	2.570	22.234
Google	2	2.708	19.93
AltaVista	2	2.754	19.96
AltaVista-2001	3	3.137	21.08
Gnutella	4	5.163	29.08
eDonkey	6	6.404	41.45

Table 1: Query length median and means across query logs.

This disparity demonstrates the differences in intent for the various types of engines. Auto-MSN queries are typically destinations, such as the hostname of a URL (e.g., *yahoo.com* and *www.hotmail.com*) that the browser doesn't recognize as a URL and instead searches. eDonkey queries, in contrast, are almost exclusively filenames (e.g., *dead can dance - arabian gothic.mp3* and *christmas songs- let it snow - frank sinatra.mp3*) that seem likely to be generated from an automated source, and are much longer. Gnutella queries appear to be human generated and describe a wider variety of content (e.g., *the last samurai*, *simcity linux*, and *john wayne*), and thus have a term-length distribution that looks more like that of Google, albeit with a smaller peak and heavier tail. See Table 2 for the most frequent queries from Google, Auto-MSN, eDonkey, and Gnutella.

The distributions and CDFs of Web query lengths are shown in Figure 2. In Figure 3 we show the query length distributions and CDFs for the two peer-to-peer traces, plus Google for reference. Figure 4 displays the relatively Zipfian rank-frequency distributions for both peer-to-peer traces plus Google.

In Figure 5, we show the query length distributions and CDFs for the three non-typical Web query traces, plus Google for reference. We see that queries to Google Groups are somewhat more likely to be single terms than

Source	Top-10 Queries
Auto-MSN	internet explorer, yahoo, google, yahoo.com.tw, www.yahoo.com.tw, www.hotmail.com, galileo.edu, www.galileo.edu, -dontrunold, www.yahoo.com
Google	lowongan kerja, beasiswa, radiokampus, sex, sms gratis, free sms, mp3, wallpaper, bugil, yahoo
Gnutella	asien porn, porn, huit femmes divx f, ebony, the last samurai, dvd, return of the king, pay it forward avi, mystic river, love actually
eDonkey	farid el, btih, dead can dance & bjork - mix.mp3, dead can dance - arabian gothic.mp3, medieval - dead can dance - celtic - wiccan - chant of the paladin.mp3, bill withers - lean on me.mp3, k1.jpg, tattoo.jpg, dead can dance - enigma of the absolute.mp3, billy ocean - suddenly.mp3

Table 2: Top-10 queries, in order, per query target.

the standard Google trace. More interestingly, queries to the Auto-MSN service are very likely to be in the form of a single term (>82%), and queries to AllTheWeb's FTP search service are comprised almost exclusively of singletons (>98.5%).

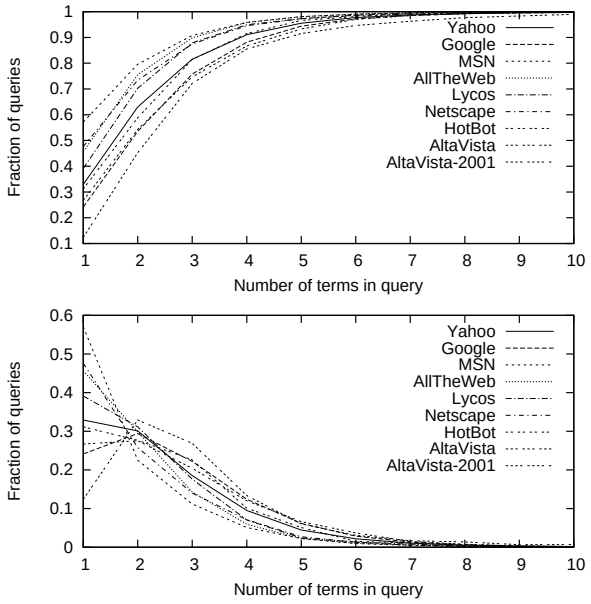
4.2 Query and term popularities

In Figure 6, we graph the frequency-rank distributions of queries and terms in the various Web traces. We also show the top ten query terms across the Web engine logs in Table 3. Most of the terms found are fairly predictable, especially when the international audience is considered. Some, however, are surprising, such as the prevalence of the term *indonesia*. While not typical of english-language searching patterns, it does reflect a significant source of queries such as *bursa musik indonesia* and *universitas indonesia*.

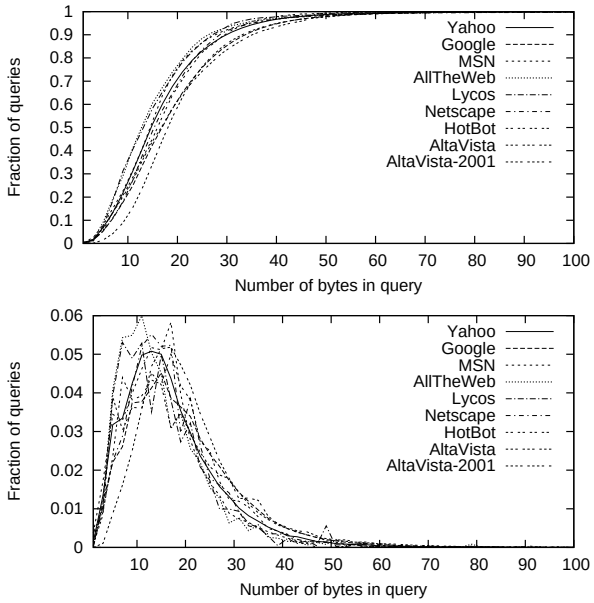
4.3 Advanced query usage

In Figure 7, we summarize the use of advanced queries. We find that, across the board, very few (less than one percent) queries utilized any of the +,-,(,) Boolean operators. Our measure likely undercounts as it measures operators common across the major search engines, but does not attempt to count explicit operators written in natural language such as OR (supported by Google) as not all engines support them, and it is often difficult to determine the original intent of the query author.

Phrasal queries, on the other hand, are relatively common, although this varies more by trace. For example, Yahoo phrase usage was approximately

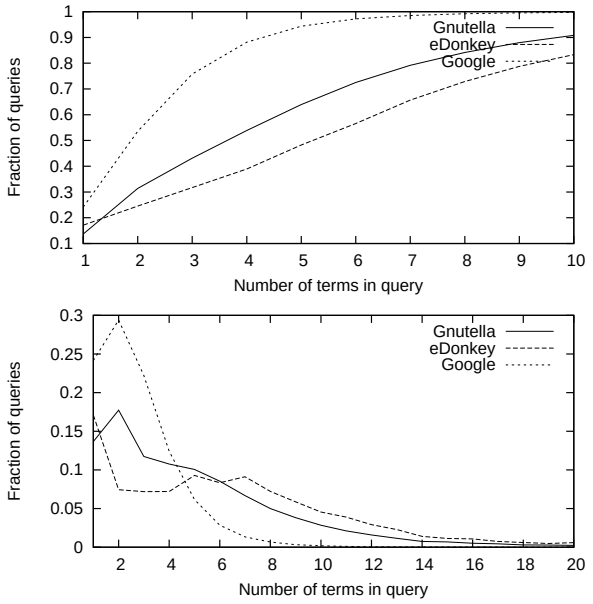


(a) Per term

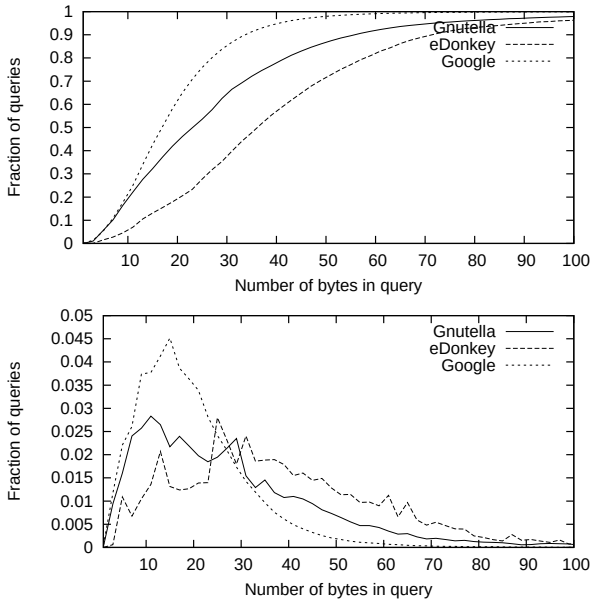


(b) Per byte

Figure 2: Web query length distributions across engines.

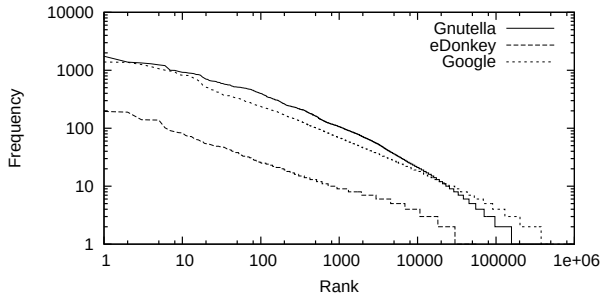


(a) Per term

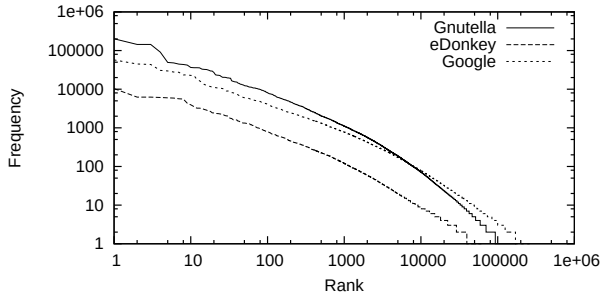


(b) Per byte

Figure 3: P2P query length distributions.



(a) queries

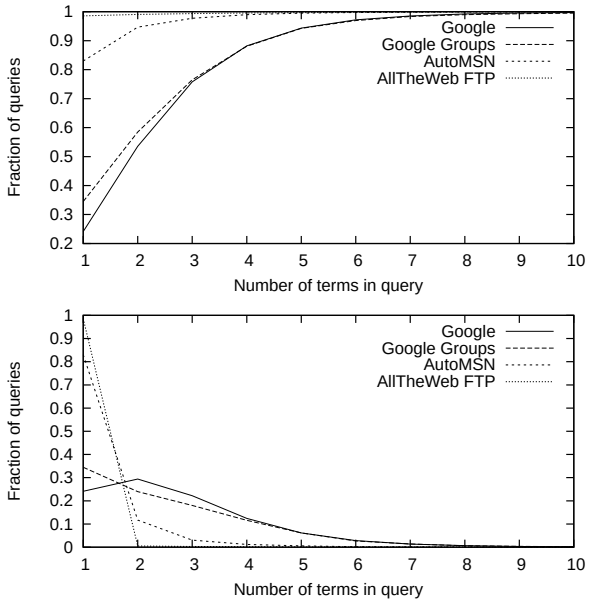


(b) terms

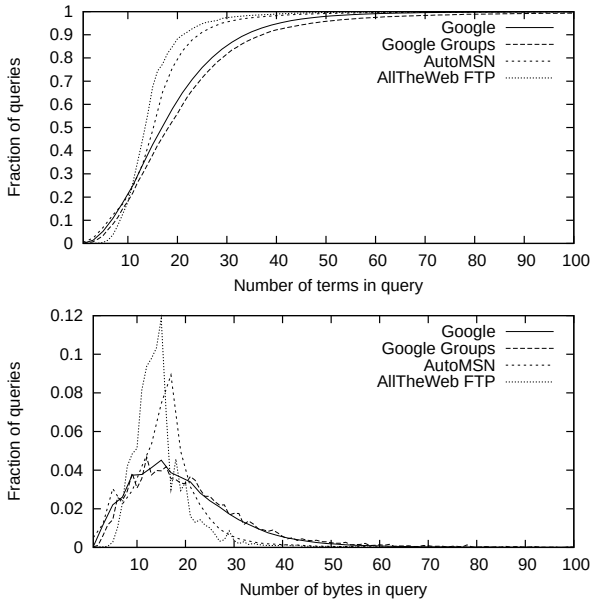
Figure 4: P2P rank-frequency distributions.

Engine	**Top-10 query terms**
AllTheWeb	free, of, nude, and, television, sex, download, de, smallville, radiokampus
AltaVista	de, la, que, es, en, of, internet, free, edge, guatemala
Google	de, of, download, free, indonesia, in, and, the, mp3, la
HotBot	and, de, for, of, the, free, in, to, a, linux
Lycos	lowongan, sapphire, of, free, application, indonesia, television, substrate, ic, 7447
MSN	of, and, internet, yahoo, explorer, in, iowa, the, sex, free
Auto-MSN	internet, explorer, yahoo, google, de, yahoo.com.tw, www.yahoo.com.tw, www.hotmail.com, galileo.edu, com
Netscape	bugil, gadis, free, >, smu, www.yahoo.com, monika, agnes, mp3search.ms.itb.ac.id, of
Yahoo	of, free, indonesia, in, and, the, sex, for, download, de
AV-2001	de, the, of, com, www, in, and, free, i, a

Table 3: Top-10 query terms, in order, per target engine.

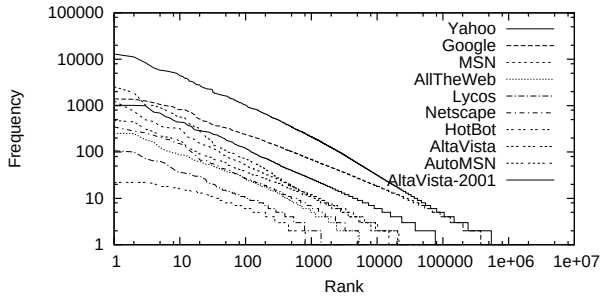


(a) Per term

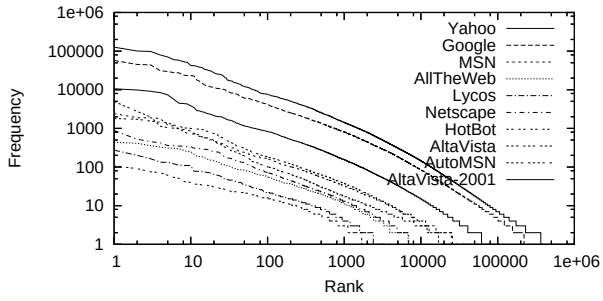


(b) Per byte

Figure 5: Query length distributions across query logs.



(a) queries



(b) terms

Figure 6: Rank-frequency distributions.

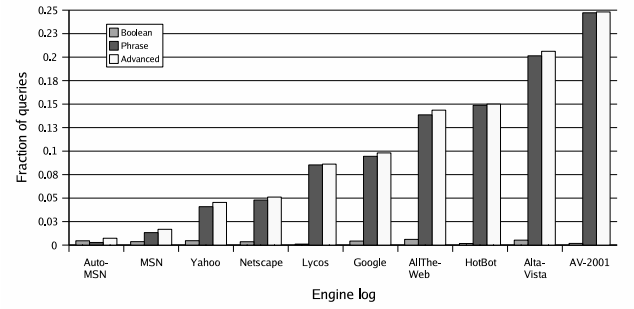


Figure 7: Boolean and phrase usage in query logs. Advanced query usage includes either Boolean or phrase query usage.

4%, while usage in AltaVista queries ranged from 20-25% (depending on trace).

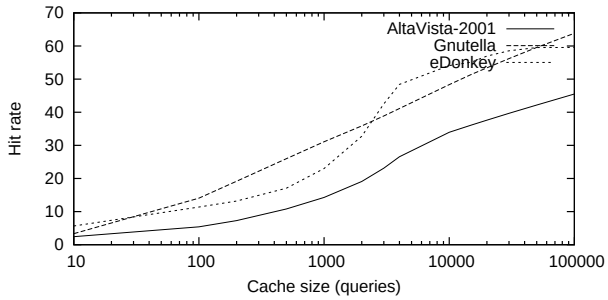
In eDonkey, an advanced queries interface allow for Boolean operations and inclusion of meta-information (such as file-type). Unfortunately, usage of advanced queries is still quite low, at just .64% of all queries recorded. Gnutella does not support Boolean operators or phrase searching.

4.4 Temporal locality

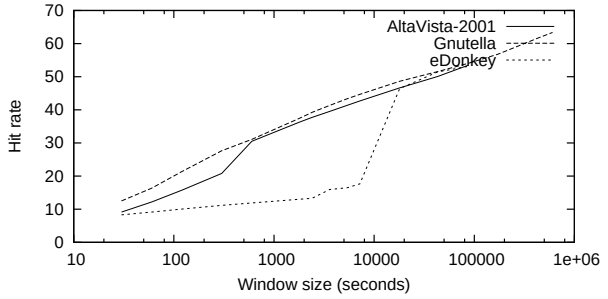
We also examined the recurrence rates of the peer-to-peer and AltaVista-2001 traces. We applied two simple caching simulators to estimate hit rates: one uses a fixed cache size and LRU replacement policy; the other a fixed window of time in which all queries received during that time are assumed cached. These are graphed in Figure 8. Caching is able to exploit significant temporal locality in the query stream. While a pure P2P cache may not be willing to cache search results for long periods [18] (as the source may no longer be online), a Web gateway (e.g., [6]) may fully exploit the query repetition by caching WWW search results for long periods.

In Figure 9(a) we plot the distribution of the distances between recurring queries in the AltaVista-2001 trace. We note the presence of a spike in activity at 300 seconds — this distinct pattern within an otherwise smooth distribution suggests automated activity every five minutes.

The distribution of distances in the eDonkey trace (shown in Figure 9(b)) also demonstrates some spikes of activity, corresponding to query repetitions at approximately 45 minutes, 1.5 hours, and 2 hours, 15 minutes. Upon examination of such queries, it is apparent that almost all of those queries are for particular filenames, suggesting perhaps an index server verifying

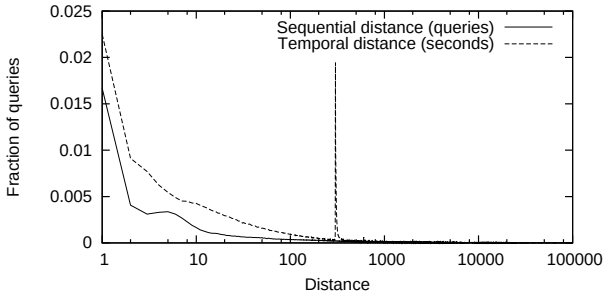


(a) LRU Cache

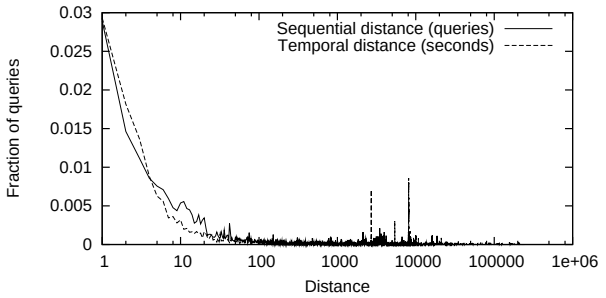


(b) Window

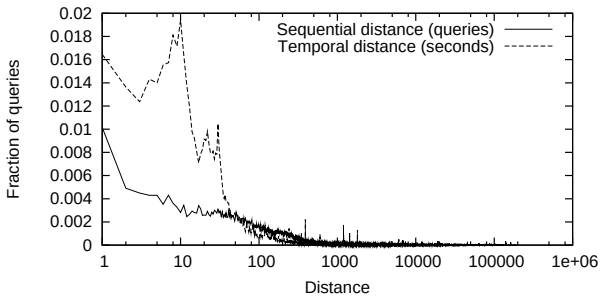
Figure 8: Caching hit rate as cache size increases.



(a) AltaVista-2001



(b) eDonkey



(c) Gnutella

Figure 9: Distributions of distances between recurring queries.

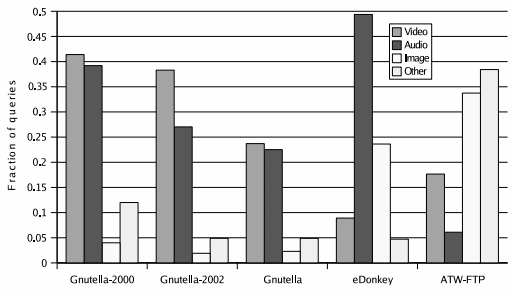


Figure 10: Types of files requested in file-oriented query logs.

that the files are still accessible.

Finally, the Gnutella distance distributions in Figure 9(c) also demonstrate unusual recurrences, most prominently around the 10 second mark.

4.5 File-type distributions

Ideally, we'd like to be able to automatically characterize the topics searched in a log. However, many languages are represented in the queries captured, making topic determinations impossible without assistance from translators.

In lieu of a semantic characterization, we performed a syntactic one, in which we examine each query for the presence of a file type extension (e.g., .exe or .mpg) at the end of a term. For comparison purposes we cluster the various file types into four categories: video, audio, images, and other (such as software and word processing documents). In Figure 10, we show the relative prevalence of each file-type category across the logs that are oriented toward file-retrieval. For comparison, we used the numbers reported by Clip2 DSS [3] for Gnutella queries in 2000, and also took the top-twenty file-types reported from a study on Gnutella [30], and labeled them under our categories. While not nearly as prevalent, we additionally show the relative frequencies of such queries in the Web search engine traces in Figure 11.

4.6 Click-throughs

The IRCACHE logs also provided a limited set of click-throughs. In particular, we found click-through logs for subsets of the AltaVista and Google logs. The Google requests include what appears to be a session or query identifier. Under that assumption, we are able to see the distribution of the number of (non-zero) click-throughs that occurred. Similarly, we used

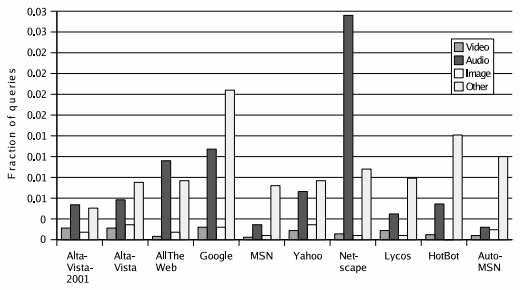


Figure 11: Types of files requested in Web search engine logs.

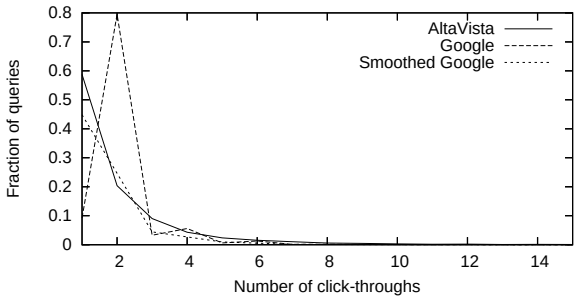
the click-through requests for AltaVista that included the query (with the unfortunate assumption that the query was not repeated and clicked-on by someone else). We found, on average, 2.07 click-throughs per AltaVista query, from a dataset of 15716 click-throughs that had queries attached. Similarly, for Google, we found 2.16 click-throughs on average per Google query, from a dataset of 68563 click-throughs with queries.

The distributions of click-throughs per query for both logs are shown in Figure 12. While the peaks at one and two clicks for AltaVista and Google, respectively (visible in Figure 12(a)), might indicate that the typical users of Google and AltaVista act differently, the presence of peaks at 2, 4, 6 (visible in Figure 12(b)), suggest at least the possibility of automated clients. When the Google click-through data is smoothed (by averaging every two points), the curve looks substantially closer to the AltaVista data.

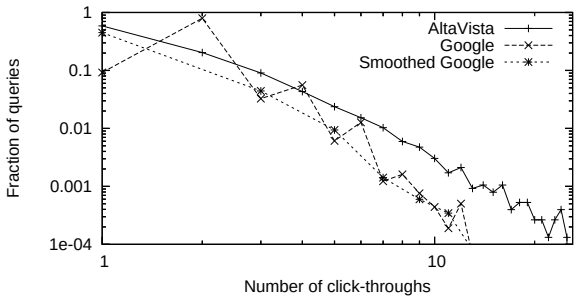
5 Summary

In this paper we have examined characteristics of a number of query logs, including queries to Web search engines, other Web services, and peer-to-peer systems. We found:

- A surprisingly low usage (less than one percent) of Boolean query operators in all logs;
- Wide variation across engines in the use of phrases (from a few percent to as much as a quarter of all queries);
- Relatively long queries for peer-to-peer traffic (median lengths of 4 and 6 versus the typical 2 for the Web);



(a) Linear scale plot.



(b) Log-log scale plot.

Figure 12: Distribution of click-throughs per query.

- Quite short queries for certain kinds of web services (FTP search and Auto-MSN service);
- On average, searchers click on slightly more than two results per query (assuming they click on anything at all);
- A significantly larger fraction of queries include file-type extensions in the peer-to-peer and FTP logs;
- A large fraction of queries in WWW and P2P logs are repeated within a short period, and a WWW search gateway can cache query results for even longer periods.

We also discovered recurring traffic patterns in the AltaVista-2001 trace and in the eDonkey and Gnutella traces, pointing to automated behavior.

In our examination of query traces from a variety of sources, we have shown relative commonality in rank-frequency distributions and the caching benefits of fixed-size caches. In addition, characteristics within each category of query logs (e.g., WWW search engines, P2P queries, and atypical Web query logs) are relatively stable. In contrast, query lengths differ considerably between Web search engines and P2P query traces, and phrase usage varies among search engines.

Acknowledgments

We thank researchers at AltaVista and Duane Wessels of Packet Pushers for providing access to the IRCACHE Web logs. We also acknowledge the support that the National Science Foundation has provided for NLANR's IRCACHE project under grants NCR-9616602 and NCR-9521745.

This material is based in part upon work supported by the National Science Foundation under Grant Numbers ANI-9903052 and IIS-0328825. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 407–415, 2000.

[2] F. Cacheda and A. Vina. Experiencies retrieving information in the world wide web. In *Proceedings of the 6th IEEE Symposium on Computers and Communications*, pages 72–79, Hammamet, Tunisia, July 2001.

[3] Clip2 DSS Group. Gnutella: To the bandwidth barrier and beyond, Nov. 2000. Available via archive.org at http://web.archive.org/web/20011212111741/http://www.clip2.com/gnutella.html.

[4] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, May 2002.

[5] B. D. Davison. A Web caching primer. *IEEE Internet Computing*, 5(4):38–45, July/August 2001.

[6] B. D. Davison, W. Zhang, and B. Wu. Lessons from a Gnutella-Web gateway. In *Alternate Track Papers and Posters Proceedings of the 13th International World Wide Web Conference*, pages 502–503, New York City, May 2004. ACM Press.

[7] B. J. Jansen and U. W. Pooch. Web user studies: A review and framework for future work. *Journal of the American Society of Information Science*, 52(3):235–246, 2001.

[8] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.

[9] B. J. Jansen, A. Spink, and T. Saracevic. Failure analysis in query construction: Data and analysis from a large sample of Web queries. In *ACM Digital Libraries*, pages 289–290, 1998.

[10] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[11] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, Edmunton, Alberta, CA, 2002. ACM.

[12] G. Kan. Gnutella. In A. Oram, editor, *Peer to Peer: Harnessing the Benefits of Disruptive Technologies*, chapter 8, pages 94–122. O'Reilly, Sebastopol, CA, Mar. 2001.

[13] B. Krishnamurthy, J. Wang, and Y. Xie. Early measurements of a cluster-based architecture for p2p systems. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, San Francisco, Nov. 2001.

[14] N. Leibowitz, M. Ripeanu, and A. Wierzbicki. Deconstructing the kazaa network. In *The Third IEEE Workshop on Internet Applications (WIAPP'03)*, San Jose, June 2003.

[15] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, May 2003.

[16] C. X. Ling, J. Gao, H. Zhang, W. Qian, and H. Zhang. Mining generalized query patterns from web logs. In *HICSS*, 2001.

[17] E. P. Markatos. On caching search engine query results. In *Proceedings of the Fifth International Web Caching and Content Delivery Workshop (WCW'00)*, Lisbon, Portugal, May 2000.

[18] E. P. Markatos. Tracing a large-scale peer-to-peer system: an hour in the life of Gnutella. In *Second IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2002.

[19] National Laboratory for Applied Network Research. A distributed testbed for national information provisioning. Home page: `http://www.ircache.net/`, 2004.

[20] P. C. Saraiva, E. S. de Moura, N. Ziviani, W. Meira, R. Fonseca, and C. Ribeiro-Neto. Rank-preserving two-level caching for scalable search engines. In *Proceedings of ACM SIGIR*, pages 51–58, New Orleans, LA, Sept. 2001.

[21] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(3), 1999. Previously available as DEC SRC Technical Note 1998-014, October, 1998.

[22] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen. U.S. versus European Web searching trends. *SIGIR Forum*, 37(2), 2002.

[23] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 53(2):226–234, 2001.

[24] K. Sripanidkulchai. The popularity of gnutella queries and its implications on scalability. Available from http://www-2.cs.cmu.edu/~kunwadee/research/p2p/gnutella.html, 2001.

[25] D. Sullivan. comScore Media Metrix search engine ratings. From Search Engine Watch, at `http://www.searchenginewatch.com/reports/article.php/2156431`, Oct. 2003.

[26] D. Sullivan. Searches per day. From Search Engine Watch, at `http://www.searchenginewatch.com/reports/article.php/2156461`, Feb. 2003.

[27] G. Wearden. eDonkey pulls ahead in European P2P race. *CNet News.com*, Oct. 2003. http://news.com.com/2100-1025-5091230.html.

[28] D. Wessels. *Web Caching*. O'Reilly & Associates, 2001.

[29] Y. Xie and D. O'Hallaron. Locality in search engine queries and its implications for caching. In *Proceedings of IEEE INFOCOM*, 2002.

[30] D. Zeinalipour-Yazti and T. Folias. A quantitative analysis of the Gnutella network traffic. Available from http://www.cs.ucr.edu/~csyiazti/courses/cs204/project/gnuDC.pdf, 2002.