Analysis of Query Logs in Gnutella Peer-to-Peer Network

ABSTRACT

Many studies focus on Web queries but few focus on peer-to-peer file-sharing system queries despite its massive scale. We analyzed several million queries collected on the Gnutella network and differentiated our findings from those of Web queries.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

Keywords

Query log analysis, Information retrieval, Peer-to-peer.

1. INTRODUCTION

Knowledge of user search patterns on a search system can be used to improve search performance. As such, many studies exist on user query logs, most of which are for the Web (e.g., [1][2]). Recently, some simulations of peer-to-peer (P2P) systems such as [4] assume that P2P queries are similar to Web queries. This is not necessarily the case. We examine a multi-million P2P query log and highlight the differences between it and Web query logs.

2. QUERY LOG ANALYSIS

To create the query log, we modified LimeWire's popular open-source implementation of the Gnutella protocol [5] to mimic a peer that can satisfy all user queries. As a result, a representative set of queries is routed to our peer and recorded.

The query log used in this study was collected during a month-long span from Sept. 14th to Oct. 14th, 2006 and includes query terms, the desired file type, and query timestamp. The queries are preprocessed by ignoring case difference, removing stop words, and replacing any punctuation with white space.

This query log (as well as others) will be published on the Web. As they are collected from a public network, they should be freely usable for research purposes. As well, the tools used to collect and analyze these logs will be made available in source form to allow the research community to either verify our findings or conduct their own analyses.

2.1 Overall Statistics

There are more than 23 million queries in our data set, of which 47% are distinct. We consider queries with an identical set of terms as similar (i.e. not distinct), even if the terms are in a different order, or have different frequencies. The total number of unique query terms (i.e. the vocabulary size) is about two million. The average query length is 3.68, which is one term longer than Web queries as popularly reported (e.g., [1]). The statistics for each query type (e.g., video, audio) are summarized in Table 1.

Table 1.General Query Statistics.

	Volume	Unique Volume	Vocaab Size	Avg. Length
No constraints	15,552,645	7,811,380	1,565,763	3.86
Audio	6,000,273	3,295,648	910,950	3.18
Video	1,439,272	598,578	235,843	2.52
Image	180,547	100,992	69,206	1.99
Document	49,196	33,976	31,248	2.22
Application	141,227	95,899	50,281	2.49
All types	23,363,160	10,762,716	2,091,464	3.68

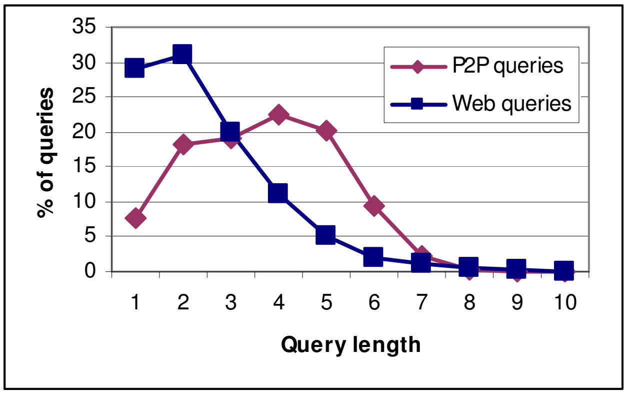

Figure 1. Query Length Distribution

As shown in Table 1, 66% of queries do not specify a type. Among the queries with specified types, more than 75% are for audio files, 20% are for video and the other 5% are for images, programs and documents. This is different from Web queries. As was reported in [2], 80% of the Web multimedia searches on Alta Vista are for images, 15% are for video, and only 5% are for audio. The reason for this phenomenon may be that there are a lot of freely downloadable, possibly illegal, audio and video files shared in P2P networks, which are not available on the Web due to copyright issues.

Query length distribution is shown in Figure 1, together with one

SIGIR'07, July 23 – 27, 2007, Amsterdam, Netherlands.

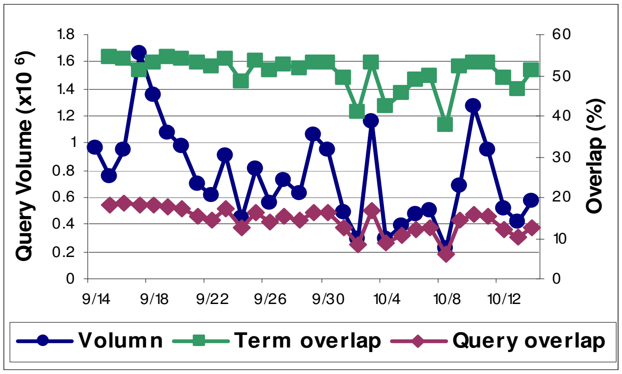

Figure 2. Query Volume and Query Overlap Over a Month.

of Web queries reported in [3]. For our data set, 80% of the queries contain from 2 to 5 terms, while only 8% of queries are single term. However, [3] reports that 76% of Web queries contain from 1 to 3 terms and single-term queries comprise as high as 28% of all the Web queries. This shows that queries in the Gnutella network tend to be longer than those in Web search. This may indicate that Gnutella users have a good idea about what they want and how to formulate a query. For example, the song title and the singer's name could be the query for searching a particular song. One such query in our query logs is "Celine Dion I am alive." More evidence of this behavior comes from the fact that audio queries are longer than all others.

2.2 Changes Over Time

User queries vary by time of day and day of a week. In Figure 2, we illustrate how query volume and content vary over a month. Query volume varies widely day-to-day, and averages about 750K per day. We also report the overlap between each day's query set and term set with those from the first day of the log. Overlap is measured by the Jaccard coefficient, defined as the ratio of the intersection of two sets to the union of the two sets. On average, there is about a 10 to 20% overlap between query sets and about a 40% to more than 50% overlap between term sets. Term overlap is naturally greater than query overlap, as the set of terms required to describe objects is more limited than their possible combinations. Furthermore, query overlap trends downward more quickly than term overlap, reflecting changing user desires.

Figure 3 shows how the set of user queries changes throughout an average day. We report the average query volume of weekends and weekdays over two weeks starting from Sept. 14. We found that on weekdays, query traffic decreases from midnight to 9am, and starts increasing after 6pm. On weekends, overall traffic volume is high during the day and only decreases after midnight. These results are different than those from Web analyses [1] and suggest that much of the usage is for recreational purposes.

2.3 Most Frequent Queries and Terms

For the set of queries, we record the most frequent queries, the most frequent query terms, and the most correlated terms for each type of queries. Due to space limits, we report below in Table 2 only the top 3 of them for audio, video, image and document types. Table 2 shows that most of video and image-constrained queries are porn-related. It also shows that some document-constrained queries are likely related to illegal activities.

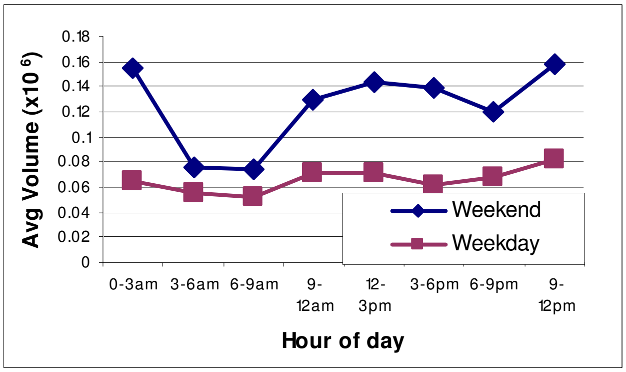

Figure 3. Query Volume of Weekdays and Weekend.

Table 2. Queries, Terms and Correlated Terms.

	Top correlated terms	**Top terms**	**Top queries**
Audio	hip hop 50 cent chain hang low	love dj remix	white and nerdy smack that chicken noodle soup
Video	nip tuck anatomy greys break prison	movie sex dvd	Pthc ptsc adult
Image	next door naar op geluk weg jan smit	qsh nude naked	Qsh ptsc lsm
Doc	Harry potter credit card ay papi	serial key music	Incest fansadox osprey

3. CONCLUSIONS

This study mainly focuses on examining a large set of queries collected from Gnutella network, with the goal of revealing the nature of queries and the users' searching behaviors. The statistical results show that P2P queries are longer than Web queries, and most searches are for music and movies. We also found that the total query traffic varies in magnitude over a month, but the level of term and query overlap is relatively stable.

In addition to user queries, we used our tools to collect information on several million files shared in the Gnutella network. Due to space limits, we omit our analyses of this data set.

4. REFERENCES

[1] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. *Hourly Analysis of a Very Large Topically Categorized Web Query Log.* SIGIR'04, July, 2004.

[2] B. J. Jansen, A. Spink, and J. Pedersen. An Analysis of Multimedia Searching on AltaVista. MIR'03, Nov, 2003.

[3] P. Reynolds and A. Vahdat. *Efficient peer-to-peer keyword searching*. ACM Conf. Middleware, 2003.

[4] J. Lu and J. Callan. *Content-based retrieval in hybrid peer-to-peer networks*. CIKM'03, Nov. 2003.

[5] T. Klingberg and R. Manfredi, *Gnutella Protocol 0.6*. rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html