Information Seeking Behavior in Peer-to-Peer Networks: An Exploratory Study

K. Y. Chan and S. H. Kwok
Department of Information and Systems Management
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{frankcky, jkwok}@ust.hk

Abstract

Digital Library (DL) has become a mean to distribute multimedia contents in the networked information environment. Previous researchers have suggested ways to improve the design of multimedia DLs, however, users' information seeking behavior is often neglected. Existing studies on users' searching behavior do not provide any insights specifically into the behavior of users searching for multimedia contents, therefore, we intend to address this issue by conducting experiments in peer-to-peer (P2P) networks, in which multimedia contents are extensively shared. Anonymity, one of the characteristics of P2P networks, has hindered previous researchers from studying individuals' information seeking behavior. In this paper, we propose a methodology to identify users' search sessions and capture their searching behavior in P2P networks at the level of individuals. Our preliminary results show that when users search for multimedia files, they are likely to submit successive queries in the same topic. The results of this study will be instrumental in providing ways of understanding users' information seeking behavior and incorporating human factors into the design of DLs.

Keywords: Digital library, information seeking behavior, peer-to-peer

1 Introduction

With technology advancements in data storage and network communications, Digital Library (DL) has become a mean to distribute multimedia contents in networked information environment. For instance, Jun [1] is an open-source graphics and multimedia library that supports multimedia contents such as movies and sound. The IRCAM Multimedia Library [8] provides an online integrated networked access to collections of multimedia contents such as audio recordings, still photos and films.

To cope with the huge volume of multimedia data, previous researchers have suggested ways to improve the design of DLs. de Vries et al. [6] developed an infrastructure for managing, indexing and serving multimedia contents in DLs. Martinez [19] proposed a library of abstract classes to model multimedia objects in a coherent, reusable and extensible framework, in order to facilitate the design and development of multimedia applications by using an object-oriented database management system. Chee et al. [2] proposed a High-speed Distributed File System (HDFS) which aimed to provide storage to large files while read and write performance would not degrade significantly compared to access for small files. Liu and Chen [17] developed a three-dimensional pattern matching mechanism for efficient video query processing. However, these works focused only on technical mechanisms of DLs, such as data organization and query processing, while they did not incorporate users' information behavior into the design of DLs.

Information seeking behavior is regarded as an important concern in the design of information systems. Information seeking behavior is defined as "the purposive seeking for information as a consequence of a need to satisfy some goal" [28]. To cope with the growing number of Internet users and the huge amount of information, designers of DLs should take account of not only the information requirements, but also users' information seeking behavior.

Prior researchers have devoted their efforts to the evaluation of user interactions with Web search engines and other information retrieval systems [4, 5, 12, 15, 21-23, 25, 26], for example, Spink et al. [25] suggested that only one in five Excite users reformulated their queries during their information-seeking processes. However, these works could only examine the searching behavior of general public as a whole, but they did not successfully look into the searching behavior of a particular group of users, who had special interests in multimedia contents, since multimedia queries accounted for only a small proportion (less than 5%) of users' web queries [24, 27]. Due to the difference in the nature between multimedia contents and other general documentary contents, users may have different information seeking behavior during their searching processes. Therefore, we choose the peer-to-peer (P2P) network, a popular platform for sharing of multimedia files, to study the information seeking behavior of users searching for multimedia contents.

The P2P also represents a simplified information-seeking environment, where extensive searching processes take place.

One of the important characteristics of P2P file-sharing network is anonymity. Existing P2P protocols [7, 9] intend to hide users' information by measures such as encrypting the query messages and not embedding locative information in the query messages. These measures that enforce anonymity have hindered the study of information seeking behavior of individuals because those query messages provide no hints on identifying the originators of a particular query message. Due to the problem of anonymity, previous researchers [13, 14] can only provide a broad picture of how a large group of P2P users behave, such as categorization of queries and ranking of popular query terms, but they provide no insight into information seeking behavior at the level of individuals.

In this paper, we propose a methodology to capture the searching activities of individuals in the Gnutella P2P network [9] and to define users' search sessions for analysis. We will also present the preliminary results of our experiment.

The organization of this paper is as follows. Section 2 defines terms used in the paper and discusses the background of Gnutella P2P networks. Section 3 introduces the proposed methodology to capture users' information seeking behavior. Section 4 presents the preliminary results of our experiment. The paper concludes in Section 5.

2 Background

2.1 Definition of terms

- A **servent** is defined as a host acting as both a server and a client in P2P networks.
- A **query** is defined as a string of key words entered by users when they perform searching.
- A **search session** is defined as the time period starting from the time the servents connect to a Gnutella network to the time the servents disconnect from the network, in which the servents have performed searching.
- A **topic** is defined as the subject to which a query belongs. We picked some topics provided in Spink, , H. C. Ozmutlu and S. Ozmutlu [26] for categorization of queries, since the P2P network contains less variety of contents than the web does. The contents offered in the web are of general and broad topics, while the contents offered in P2P networks are usually more specific to serve particular interests, such as entertainment and computer software. For example, some topics, such as News, Medical, Business and Shopping, exist in the web but are not applicable to the contents in P2P networks. P2P queries are categorized into five topics: Computer, Education, Entertainment, Sexual and Inexplicit.
  - **Computer**: Includes software, applications and games.
  - **Education**: Includes books and other documents for educational purposes.
  - **Entertainment**: Includes multimedia files for entertainment purposes, such as pop songs and movies. Our study will focus on this topic.
  - **Sexual**: Includes sexual contents, such as pornography.
  - **Inexplicit**: Includes any other unclassified queries.

  We do not further narrow down the topics because it is difficult to infer the specific purpose of a query from just a few terms. For example, a query with an actor's name may aim for a movie or an image, so we can classify the queries with broad topics only.
- **Metadata** is a definition or description of data. Examples include the title of a song and the director of a movie.

2.2 Gnutella Protocol Specification 0.4

Gnutella protocol 0.4 [3] employs a pure decentralized model [20]. In this model, all servents are equal in terms of functionality. They offer client-side functions such as accepting queries from users and returning search results, while they also perform server-side roles such as matching incoming queries against their local resources and returning results to other servents.

The Gnutella protocol defines five different types of packets, namely QUERY, HIT, PING, PONG, and
PUSH.

1. **QUERY** – requests for sharable files matching the given criteria.
2. **HIT**- responds to a query by returning a list of sharable files matching the given criteria and the IP addresses of their providers. There may be multiple responders to a query due to the nature of distributed network.
3. **PING** - requests the transitive closure of connected peers to identify themselves.
4. **PONG** - responds by a peer upon receiving a PING, the responding peer provides its IP address and the number of sharable files it contains.
5. **PUSH** - requests a file provider to contact the requester. This provides a simple mechanism to attempt to get through the firewalls.

To search for files in the Gnutella network, the Gnutella servent sends a QUERY packet that contains the query string. The neighboring servents

will receive this QUERY packet and they will take the query string as search criteria to check their sharable local files that match the criteria. If the servent finds relevant files in its local resource, it will generate one or more HIT packets containing information about the relevant sharable files. The HIT packets will be back-propagated, i.e., sent on the reverse of the path taken by the initial QUERY message, to the requesting servents.

For the searching operation, messages are distributed based on a flooding mechanism, in which a servent receiving a query message will make copies of the query message and broadcast them to all its neighboring servents. A mechanism named time-to-live (TTL) counter is introduced in Gnutella protocol in order to relieve the excessive network traffic caused by the message flooding mechanism. TTL counter, which default value is 7, is decreased by one when the query message passes through a peer. The query messages are relayed until their embedded TTL counters reach zero.

To monitor the searching activities in the Gnutella network, previous researchers [13, 14] placed a monitoring servent with log-recording capability to capture all the query messages which passed through the monitoring servent. This approach could capture the query terms searched by the users as a whole, however, it failed to identify the originators of query messages and to keep track of the searching activities of individuals. Thus, search sessions of individuals cannot be identified. Figure 1 illustrates the topology of the experiment.

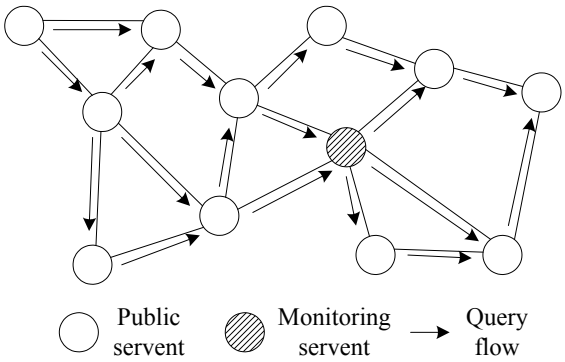

Figure 1: Topology of the Gnutella (Protocol Specification 0.4) Network.

2.3 Gnutella Protocol Specification 0.6

Gnutella protocol 0.6 [10] employs a hybrid architecture [20] combining centralized and decentralized model. Servents are categorized into "leaf" and "ultrapeer". To be qualified as utlrapeers, servents should possess sufficient computational power, network bandwidth and long expected uptimes. Detailed principles of electing ultrapeers are described in the Gnutella Protocol Specification v0.6 [10].

An ultrapeer maintains connections with other ultrapeers and acts as a proxy to the Gnutella network for the leaves connected to it. A leaf keeps only a small number of connections to ultrapeers. An ultrapeer only forwards a query to a leaf if it believes the leaf can answer it, and leaves never relay queries between ultrapeers. Figure 2 illustrates the topology of the Gnutella (Protocol Specification 0.6) network.

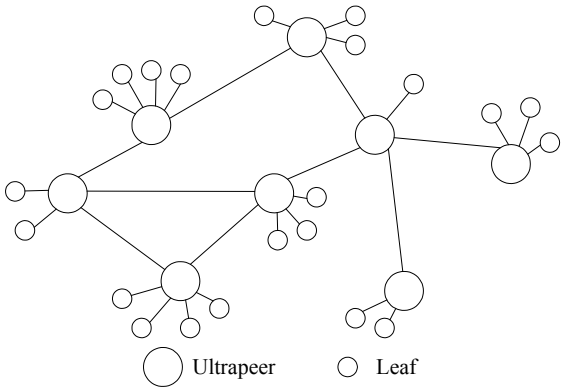

Figure 2: Topology of the Gnutella (Protocol Specification 0.6) Network.

In this model, since an ultrapeer acts as a proxy to the Gnutella network for it connected leaves, not only can it collect the query messages at the ultrapeer level through the ultrapeer-to-ultrapeer connections, but also it can keep track of searching activity of individual leaf through the ultrapeer-to-leaf connections. By maintaining a number of leaf connections, an ultrapeer can uniquely identify the originators of query messages sent by its connected leaves. For instance, it is possible to record the searching activity of a leaf throughout the period starting from the time the leaf connects to the time the leaf disconnects. Tracking of searching habit of individual P2P users becomes possible by implementing a monitoring ultrapeer in the Gnutella network.

3 Research Design

In related research, Markatos [18] has suggested that the overall P2P network characteristics can be represented by a randomly-chosen peer in the P2P network. Not only does every peer have similar network characteristics, but also the characteristics are location independent. To capture users' information seeking behavior in the P2P network, a Gnutella P2P servent was developed based on an open-source servent, LimeWire [16]. The P2P servent was connected to the Gnutella P2P network and was programmed to always act as an ultrapeer. The servent accepted incoming connections from leaves and routed the queries from its shielded leaves to other connected ultrapeers. Therefore, the

servent was able to capture the searching activities of its shielded leaves. Figure 3 illustrates the network connections between our servent and other P2P servents.

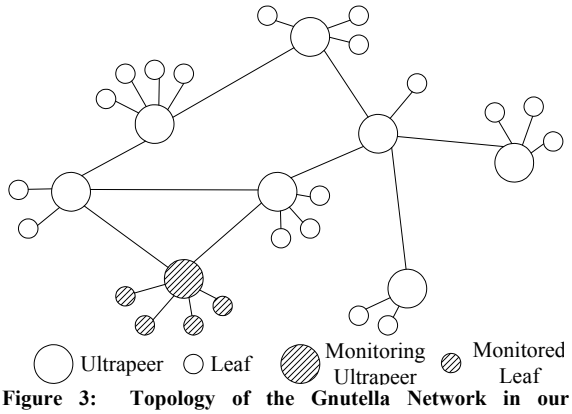

Figure 3: Topology of the Gnutella Network in our experiment.

Our Gnutella servent operated on a broadband connection through the Internet Service Provider (ISP), i-Cable, in Hong Kong. A set of data was collected from 0:00, 3 October 2003 to 0:00, 4 October 2003 Hong Kong time. To improve the efficiency of data collection, our servent was programmed to maintain at most 20 ultrapeer connections and 100 leaf-node connections simultaneously. There were 3025 servents connecting to our servent as leaves in the one-day data collection period. The searching activities of these servents were recorded in a log file, which was later imported into a database system for processing. Table 1 summarizes the settings of our data collection experiment. Table 2 shows an example of the log file recording queries from an individual servent.

Table 1: Summary of our experiments.

Date	Duration	Number of ultrapeer connections (maximum)	Number of leaf connections (maximum)	Number of monitored hosts
October 3, 2003	24 hours	20	100	3025

Table 2: An example of log file.

Time	Host IP	Search Criteria/ Connection Status	GUID
9:16:50	a.b.c.d	***connected***	
9:22:11	a.b.c.d	Country	74FF168526EFE5AF FF144527D2ED9500
9:26:09	a.b.c.d	im gonna take the mountain	09964821B6CFDF1C FF808D7794838600
9:26:30	a.b.c.d	cowboys like us	0B4E54F0197DC63B FF73379CAB3C0700
9:28:07	a.b.c.d	perfect	AA29CAEE72FF1FCB FFE91CBE8BAA4500
9:30:04	a.b.c.d	senorita	76CE25D54178C320F F6365EA9A0DBD00
9:35:54	a.b.c.d	***disconnected***	

***Host IP** is hided for privacy reasons.

4 Results

4.1 Analysis on queries in terms of search topics

The queries are categorized into five categories: Computer, Education, Entertainment, Sexual and Inexplicit. Each query is manually entered in the Google web search engine [11] to verify its belonging category. The categorization process is based on the returned results from the search engine. For example, if the majority of returned results is related to entertainment, the query will be categorized into the topic named Entertainment. Moreover, the file types of desired contents, such as Audio and Video, help us categorize the queries. For example, if the query is specified with the file type named Audio and the search criterion is a singer's name, the query is classified to be of the topic named Entertainment. The distribution of queries is shown in Table 3. It is not surprising that the most popular category is Entertainment, which accounts for about 53% of queries in the sample, since P2P network is mainly used for sharing multimedia files for entertainment purpose. This figure also ratifies our claim that the P2P network is a suitable platform to capture the information seeking behavior of users searching for multimedia contents. About 12% of queries are classified as Inexplicit because some users submit queries of one to two terms, such as "nice" and "down", which provide no indication of what they search for.

Table 3: Distribution of number of queries to topics.

Topic	Number of Queries	%
Computer	488	10.59%
Education	51	1.11%
Entertainment	2442	52.98%
Sexual	1062	23.04%
Inexplicit	566	12.28%
Total	4609	100.00%

4.2 Analysis on queries in terms of session

Among 3219 connection sessions from 3025 servents, 661 were search sessions which included one or more than one query. The mean number of queries per session for the Gnutella P2P network is 6.97. Figure 4 shows the distribution for number of queries per session. It is shown that most sessions tended to have a small number of queries per session.

4.3 Analysis on queries in terms of topic change

There were 256 topic changes in 661 search sessions, whereas the mean number of topics per session is 1.3 and the mean number of topic changes per session is 0.4. It means that in a search session,

an average user searched for files on one topic only and seldom switched to other topics. Figure 5 shows the distribution for number of queries per topic. As it is observed, usually 1 to 5 consecutive queries were spent on a particular topic. On the other hand, 14.5% of users submitted more than 30 consecutive queries on a single topic. The mean number of queries per topic is 17.8 queries, which means that users searched for another topic every 17.8 queries. It should be noted that this figure is deviated by the significant amount of users who submitted more than 30 successive queries on a single topic.

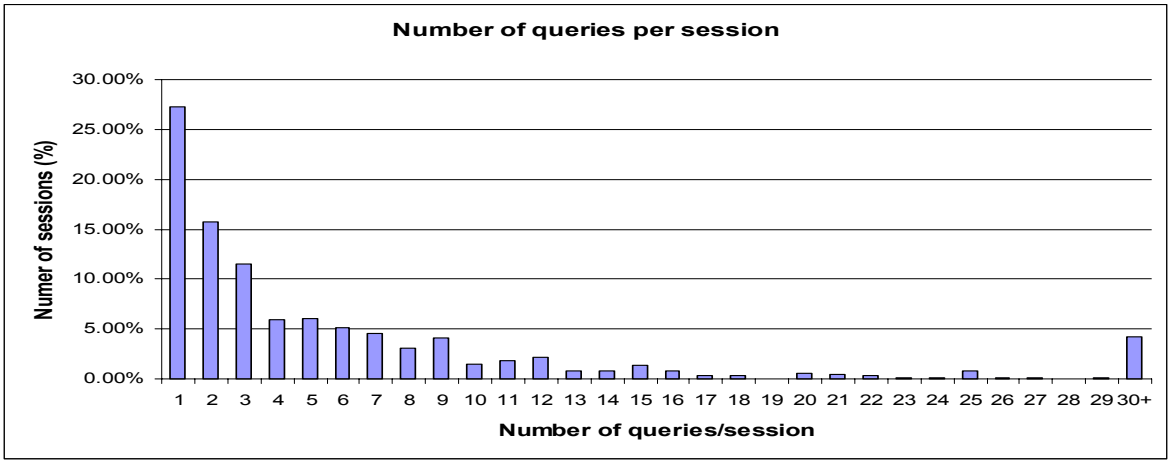

Figure 4: Histogram for number of queries per session.

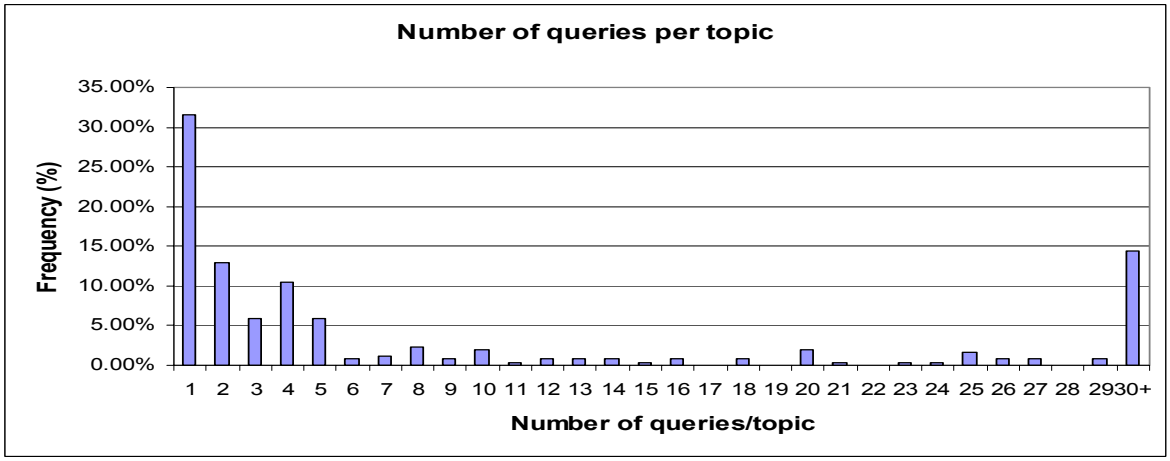

Figure 5: Histogram for number of queries per topic.

4.4 Analysis on queries in terms of file type specification

In our experimental environment, users can voluntarily specify the file types of their desired contents when they submit queries, a feature aiming to improve the accuracy of searching. The file types include Audio, Documents, Images, Software Programs, Video and Any types. Among 4609 queries, only 1367 queries are specified with file types, while about 70% of queries are not specified. Among those file types, only Audio and Video are used by users to specify their queries. Table 4 shows the distribution of queries specified with file types among different topics. About 46% of users tend to specify the file types, Audio or Video, when they look for multimedia files for entertainment purposes.

Table 4: Distribution of queries specified with file types among different topics

Topic	Specified with file types	Unspecified	Total queries
Computer	0 (0%)	488 (100%)	488
Education	0 (0%)	51 (100%)	51
Entertainment	1127 (46%)	1315 (54%)	2442
Sexual	213 (20%)	849 (80%)	1062
Inexplicit	27 (5%)	539 (95%)	566

4.5 Analysis on queries in terms of metadata search

To search for contents in P2P networks, users can include metadata in their queries to narrow the scope of searching. In our experiment, for example, users can specific their queries with Title, Artist, Album

and Genre to locate audio contents, while they can specific their queries with Title, Type, Year and Director to locate video contents. Among 994 queries searching for audio contents, 82.09% of queries include 1 metadata field, 17.3% of queries include 2 metadata fields, while 0.6% of queries include 3 metadata fields. Among those 373 queries searching for video contents, 91.96% of queries include 1 metadata field, 7.51% queries include 2 metadata fields and 0.54% of queries include 3 metadata fields. For both audio and video queries, none of them includes all 4 metadata fields. Table 5 summarizes the distribution of queries with different number of metadata fields. As it is observed, users are likely to include 1 metadata field when they perform metadata search for multimedia contents. They seldom specify their queries with more than one metadata field. Table 6 presents the distribution of queries including different metadata fields. Title and Artist are the popular fields when users perform metadata search for audio contents, while Title is the most popular field for metadata search for video contents. These figures suggest that those metadata fields are not of equal interest among users. Some metadata fields are used more often than others are.

Table 5: Distribution of queries with different number of fields of metadata

Type \ Number of fields	1	2	3	4
Audio	82.09%	17.30%	0.60%	0.00%
Video	91.96%	7.51%	0.54%	0.00%

Table 6: Distribution of queries including different metadata fields

Type \ Metadata field	Title	Artist	Album	Genre
Audio	47.89%	57.24%	10.36%	2.62%
Video	94.64%	11.26%	1.07%	1.07%

4.6 Analysis on queries in terms of duration of sessions

The mean duration of the 661 search sessions is 3087.9 seconds, whereas the standard deviation is 3707.1 seconds. This shows a significant variation in the duration of search sessions in P2P network. This may be due to the dynamic nature of P2P network, in which servents frequently disconnect from the P2P network. The distribution of session durations is shown in Figure 6. A major proportion of sessions were under 2000 seconds. Only 2.28% of sessions lasted for more than 12000 seconds. Table 7 shows the summary of our data collected in the Gnutella P2P network.

Table 7: Summary of data collected in the Gnutella P2P network.

Total number of search sessions	661
Mean of duration of search session (sec)	3087.9
S.D of duration of search session (sec)	3707.1
Total number of queries	4609
Mean of queries per session	6.97
Total number of topic change	256
Mean of topics per session	1.3
Mean of topic changes per session	0.4
Mean of queries per topic	17.8

5 Conclusions and Future Work

In this paper, we introduced a methodology to capture the information seeking behavior of users searching for multimedia contents, and preliminary results were presented. The contributions of this paper are (1) to help understand users' searching behavior in P2P networks at the level of individuals, while previous works could only capture the

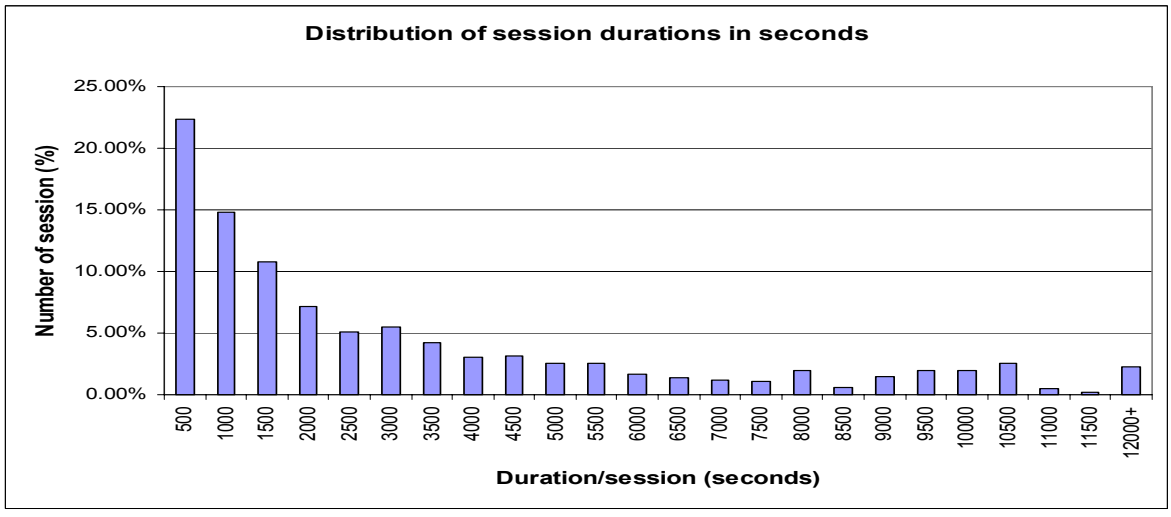

Figure 6: Histogram for distribution of session duration.

behavior of general public as a whole; and (2) to study, through experiments, the information seeking behavior of real users in Gnutella P2P networks, in which multimedia files were extensively shared. The findings make contributions to the design of multimedia DLs.

DLs can incorporate some features to accommodate users' information seeking behavior. For example, when the DL receives queries about entertainment, the system should be able to predict what the users are looking for, since users seldom change their searched topic within a single search session, as observed from our preliminary results. The DL can refer users' queries and its predictions of users' requests to other DLs in advance, if it does not contain the requested contents, to improve its response time. Moreover, since not all metadata fields are not considered useful, metadata fields should be carefully chosen to tailor users' searching needs. The understanding of users' information seeking behavior also helps address individual needs and provide users with related contents of their interests, in order to facilitate push of information in DLs. Finally, organizing data according to users' searching behavior may be a way to shorten information retrieval time, which is a crucial performance indicator in DLs.

Future works of our study include the analysis of changes in users' information seeking behavior over a certain period of time and how users behave when they engage in multitasking searching.

Acknowledgments

The work described in this paper was partially supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKUST6256/03E).

References

[1] A. Aoki, K. Hayashi, K. Kishida, K. Nakakoji, Y. Nishinaka, B. Reeves, A. Takasbima, and Y. Yamamoto, "A case study of the evolution of Jun: an object-oriented open-source 3D multimedia library," presented at 23rd International Conference on Software Engineering, 2001.

[2] C. L. Chee, S. S. Erdogan, C. W. Ngo, and C. K. Wong, "A high speed distributed file system for multimedia communications," presented at 1994 IEEE Region 10's Ninth Annual International Conference, 1994.

[3] Clip2, The Gnutella Protocol Specification v0.4, http://www9.limewire.com/developer/gnutella_protocol_0.4.pdf, accessed on 23 October 2003.

[4] M. D. Cooper, "Usage patterns of a Web-based library catalog," *Journal of the American Society for Information Science & Technology*, vol. 52, pp. 137-148, 2001.

[5] V. Cothey, "A longitudinal study of World Wide Web users' information-searching behavior," *Journal of the American Society for Information Science & Technology*, vol. 53, pp. 67-78, 2002.

[6] A. P. de Vries, B. Eberman, and D. E. Kovalcin, "The design and implementation of an infrastructure for multimedia digital libraries," presented at International Database Engineering and Applications Symposium (IDEAS'98), 1998.

[7] eDonkey, eDonkey2000, http://www.edonkey2000.com/, accessed on 23 October 2003.

[8] M. Fingerhut, "The IRCAM multimedia library: a digital music library," presented at IEEE Forum on Research and Technology Advances in Digital Libraries, 1999.

[9] Gnutella, Gnutella, http://rfc-gnutella.sourceforge.net/research/index.html, accessed on 23 October 2003.

[10] Gnutella, The Gnutella Protocol Specification v0.6, http://rfc-gnutella.sourceforge.net/developer/testing/index.html, accessed on 23 October 2003.

[11] Google, Google, http://www.google.com/, accessed on 23 October 2003.

[12] C. Holscher and G. Strube, "Web search behavior of Internet experts and newbies," *Elsevier. Computer Networks: the International Journal of Distributed Informatique*, vol. 33, pp. 337-346, 2000.

[13] S. H. Kwok, S. M. Lui, R. Cheung, S. Chan, and C. C. Yang, "Searching Behavior in Peer-to-Peer Communities," presented at Proceedings of the International Conference on Information Technology: Computers and Communications (ITCC.03), 2003.

[14] S. H. Kwok and C. C. Yang, "Searching the Peer-to-Peer Networks: The Community and Their Queries," *Journal of the American Society for Information Science and Technology (JASIST)*, 2004 forthcoming.

[15] A. W. Lazonder, H. J. A. Biemans, and I. Wopereis, "Differences between novice and experienced users in searching information on the World Wide Web," *Journal of the American Society for Information Science*, vol. 51, pp. 576-581, 2000.

[16] Limewire, Limewire.org, http://www.limewire.org/, accessed on 23 October 2003.

[17] C.-C. Liu and A. L. P. Chen, "3D-List: a data structure for efficient video query processing," *IEEE Transactions on Knowledge & Data Engineering*, vol. 14, pp. 106-122, 2002.

[18] E. P. Markatos, "Tracing a large-scale peer to peer system: an hour in the life of Gnutella," *Proceedings CCGRID 2002. 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid. IEEE Comput. Soc. 2002*, pp. 65-74, 2002.

[19] J. Martinez, "The design of an extensible multimedia library for an OODBMS," presented at Seventh International Workshop on Database and Expert Systems Applications, 1996.

[20] N. Minar, Distributed Systems Topologies: Part 2, http://www.openp2p.com/pub/a/p2p/2002/01/08/p2p_topologies_pt2.html, accessed on 23 October 2003.

[21] A. Spink, "A user-centered approach to evaluating human interaction with Web search engines: an exploratory study," *Information Processing & Management*, vol. 38, pp. 401-426, 2002.

[22] A. Spink and H. Cenk Ozmultu, "Characteristics of question format Web queries: an exploratory study," *Information Processing & Management*, vol. 38, pp. 453-471, 2002.

[23] A. Spink, H. Cenk Ozmutlu, and S. Ozmutlu, "Multitasking information seeking and searching processes," *Journal of the American Society for Information Science & Technology*, vol. 53, pp. 639-652, 2002.

[24] A. Spink, A. Goodrum, and A. R. Hurson, "Multimedia Web queries: implications for design," presented at International Conference on Information Technology: Coding and Computing, 2001.

[25] A. Spink, B. J. Jensen, and H. C. Ozmultu, "Use of query reformulation and relevance feedback by Excite users," *Internet Research-Electronic Networking Applications & Policy*, vol. 10, pp. 317-328, 2000.

[26] A. Spink, H. C. Ozmutlu, and S. Ozmutlu, "A Study of Multitasking Web Search," presented at International Conference on Information Technology: Computers and Communications (ITCC'03), 2003.

[27] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen, "U.S. versus European Web searching trends," *ACM SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, vol. 36, pp. 32-38, 2002.

[28] T. D. Wilson, "Human information behavior," *Informing Science*, vol. 3, pp. 49-55, 2000.