Searching Behavior in Peer-to-Peer Communities

Sai Ho Kwok[1], S. M. Lui[1], Ricky Cheung[1], Sally Chan[1], and Christopher C. Yang[2]
[1]*Department of Information and Systems Management*
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong SAR
{jkwok, imcarrie, rickyc, sallyc}@ust.hk
[2]*Department of Systems Engineering and Engineering Management*
The Chinese University of Hong Kong
Shatin, N.T. Hong Kong SAR
yang@se.cuhk.edu.hk

Abstract

The focus of this paper is the searching behavior in Peer-to-Peer (P2P) communities and in that context, will discuss potential for enhancement of the P2P protocol using searching behavior. Due to excessive network traffic, network scalability is a major P2P problem. The purpose of this paper is to study searching behavior by using log files and subsequently, will propose solutions to reduce the amount of redundant searching activities. In the current research, log data containing over 5 million QUERY messages was collected over seven consecutive days using a Gnutella-based P2P program. The results of this study will be instrumental in providing ways to reduce existing problems with P2P network traffic.

1. Introduction

For clarification, Peer-to-Peer (P2P) is a technique that can be described as facilitating file sharing over a P2P network. Specifically, the P2P networks (or communities) contain a large number of nodes (or peers). These nodes, also known as servants, act simultaneously as both clients and servers. The popularity of P2P file sharing has attracted considerable interest. Consequently, a large number of researchers have investigated its extensibility and applicability. Under closer scrutiny, it would seem that network scalability is an inherent problem of P2P [7]. Many researchers have devoted their efforts to network analysis, for example network traffic patterns [5] and network performance measurement [10] in an attempt to find ways reducing the P2P network traffic and improving the efficiency of P2P file sharing. However, the key factor of searching behavior is how P2P users behave in the P2P network. This factor has not been explicitly addressed in previous studies and relevant data is lacking in the literature. The current paper will attempt to resolve this problem by observing peer's activities with use of the servant's log files, advocating ways to reduce network traffic.

Searching is a major component of P2P file sharing. One of the most prevalent problems with P2P use is that a considerable amount of time is spent in searching. The reason for this is that the clients usually do not find the files they need. Even when they have found the required files, they have to query the same files again for other sources when the remote peers disconnect from the network. Consequently, there is a large amount of query messages flooding the P2P network that will jeopardize the interests of the P2P communities. One of the ways to enhance the P2P network and file sharing protocol is to understand the searching behavior of P2P users. For example, it is important to know how the query messages are generated and what kind of query message most frequently occupies the P2P network. In addition, it is necessary to have access to information on the kinds of files that the P2P users query most because the searching behavior has a direct impact on the network traffic.

2. P2P Network Characteristics

Several previous research projects on related topics have collected data on P2P network characteristics. An example of these is the work of Adar and Huberman [1] who recorded the P2P network traffic for 24 hours

continuously. They have indicated that free riding is a serious problem for the P2P network. The free-riding problem refers to the tendency of many P2P users who request to download a file but rarely share their own files with others. Markatoa [5] has investigated the magnitude and traffic patterns of the P2P network by tracing the queries going through a P2P servant in an hour. In yet another related study Matei et al. [6] sent a crawler to collect the topology information of the P2P network. The study evaluated costs and benefits of the P2P approach and according to the data, the mismatching of the P2P and Internet infrastructure topology have considerable impact on the overall performance of P2P system. These studies have all discussed the network traffics and connectivity of the peers, which provide implications of the topology design for P2P application. In addition, Saroiu et al. [7] studied the characteristic of the participating peers. These findings have suggested that there is significant heterogeneity and lack of cooperation among peers participating in P2P network. However, an important aspect of the system, namely searching behavior, appears to be a missing factor in the analysis of P2P network.

3. Data Collection

In related research, Markatos [5] has shown that the overall P2P network characteristics can be represented by a randomly-chosen node in the P2P network. Not only does every P2P node have similar network characteristics but the characteristics are also location independent. In the present experiment, the query data was collected from 16 July through to 22 July 2002 (7 consecutive days) with a P2P servant situated in Hong Kong. The P2P servant ran on a P4 1.2G PC with 100Mps bandwidth. The P2P servant used for data collection was written in Java and based on the Java API of Jtella [2] that follows the Gnutella protocol [3]. All query messages going through the P2P servant were collected in a log file, the monitor log file. Two sample data in the monitor log is shown in Table 1. The monitor log file was then imported to a database management system for data analysis.

4. Query Analysis

In the Gnutella protocol, there are five kinds of messages for communications among peers and they are PING, PONG, QUERY, QUERYHIT, and PUSH. The focus of the present study is limited to the QUERY message. A QUERY message contains the query terms that are usually entered by the P2P user or those that are sometimes issued by the P2P servants. Over the seven days of continuous operation, the P2P servant for the study collected a total of 5,052,754 QUERY messages from the P2P network. These were stored in the monitor log. It has been observed that 28% of all query messages contained two terms, a figure that is very similar to the findings in web searching [4]. The distribution of the number of terms per query is presented in Figure 1.

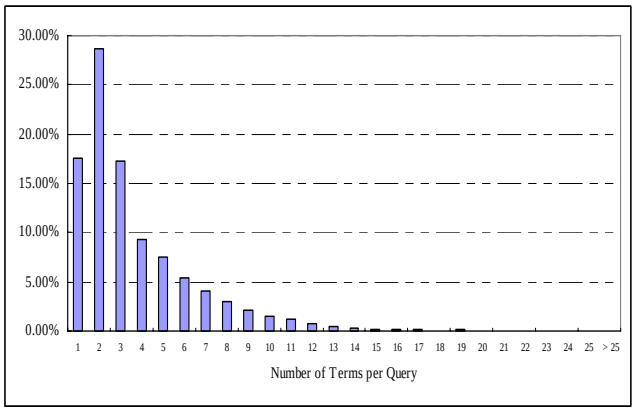

Figure 1: Number of Terms per query.

Table 1: Sample data in the monitor log.

Date - Time	Min. Speed	TTL	Hops	GUID	Origin Host:Port	Search Criteria
8/6/2002 16:08	33665	0	1	[dd]-[d1]-[a8]-[18]-[e0]-[7c]-[83]-[de]-[ff]-[d5]-[13]-[e3]-[bf]-[7a]-[aa]-[0]	connect2.gnutellanet.com:6346	dvia
8/6/2002 16:08	44417	0	1	[92]-[c0]-[bc]-[cc]-[88]-[90]-[a7]-[28]-[ff]-[2e]-[30]-[4]-[77]-[36]-[30]-[0]	connect1.gnutellanet.com:6346	good mp3

4.1. Query Types

Based on these results, it would appear that the typical query types are similar to those used in the study of Web searching log [8] with some minor modification. They are unique queries, repeat queries, zero-term queries, XML queries, and non-English queries. Definitions of these specific types are provided in Table 2. Table 3 presents the statistical results of query message used in the monitor log of the present study. The monitor log contains a record of query messages going through a P2P servant. The results indicate that 97.5 % of queries are repeating queries and that these repeat queries are repeated, on average 7.16 times. However, in extreme cases, some repeat queries repeated over 10,000 times. For example, the term "divx" repeatedly occurs, being recorded more than 10,000 times. It is useful to know that the percentage of non-English query is only less than 0.6 % and the percentage of XML query is about 0.8%. These findings would indicate that this form of query is in the minority in among P2P communities studied for the current project. Perhaps, this also indicates that English speaking P2P users are predominant in these particular communities. Moreover, the results show that only a few P2P servants support advanced XML-based queries. The findings of the present study support the fact that the zero-term queries are rare because many P2P servants disallow zero-term queries. The important figures in the statistical results are that there are over 97.5 % repeat queries and only 2.5% unique queries in the P2P network. These results provide clear evidence that the P2P network was filled with repeat queries. This implies that the network traffic can be reduced when the amount of repeat queries is decreased.

4.2. Query Content

In studying the query content, all queries were grouped and sorted according to the frequency in numbers of occurrence. Table 4 lists the top 15 queries found in the monitor log. From this list the observation can be made that P2P users are clearly interested in up to date content. Examples of this are queries related to recently released movies such as "Spiderman" or "Minority report", popular artist's names such as "Eminem", "Nelly", and "Chris Isaak" or band names such as "Aqua mp3". Another popular category of query is adult content ("Porn", "Porn mpg", and "Sex") and this matches with the behavior of web searching [8]. Moreover, some common queries refer to file extensions only ("Divx", and "Divx avi") and it is rare in web searching.

Table 2: Definitions of Query Types.

Unique queries	The queries with different combination of terms entered by a user, appeared only once throughout the experiment
Repeat queries	The queries appeared more than one throughout the experiment
Zero-term queries	The queries with a non-query term "[enter search term]", "enter search term]", "enter search here", or "*** enter search here ***".
XML queries	The queries containing a XML substring "<?xml version ="1.0"?>" and usually generated by advanced Gnutella clients; e.g., LimeWire [9]
Non-English queries	The queries containing non-English characters; e.g., Chinese

Table 3: Occurrence of different types of query.

	Occurrences	**%**
Total Number of Queries (7 days)	5052754	100%
Unique queries	128001	2.5 %
Repeat queries	4924753	97.5 %
Zero term queries	2539	<0.05 %
XML queries	39151	<0.8 %
Non-English queries	29684	<0.6%

Table 4: Top 15 queries in the P2P network.

Rank	**Query**	**Occurrence**	**%**
1	divx	11820	0.23
2	qwerty jpg	8746	0.17
3	porn	6510	0.13
4	eminem	6365	0.13
5	techno mp3	6159	0.12
6	divx avi	5385	0.11
7	porn mpg	4805	0.10
8	spiderman	4402	0.09
9	chris isaak	4370	0.09
10	return to me	4115	0.08
11	joey gian	4088	0.08
12	Nelly	3845	0.08
13	Sex	3567	0.07
14	Aqua mp3	3567	0.07
15	minority report	3330	0.07

Table 5 presents the top 15 file types specified queries. Table 6 classifies queries into different file types. A common misconception has been that P2P was dedicated to music sharing but surprisingly the most demand is for the video file type. Consequently these findings are able to explain why the Internet becomes overloaded when more P2P users try to download video files.

Table 5: Top 15 file types specified in queries.

Rank	File extension	Occurrence	%
1	mp3	1149893	22.76
2	avi	859048	17.00
3	mpg	491146	9.72
4	zip	83001	1.64
5	mpeg	81526	1.61
6	jpg	54791	1.08
7	asf	40924	0.81
8	divx	35040	0.69
9	Ra	32214	0.64
10	rm	21094	0.42
11	pdf	19792	0.39
12	exe	17369	0.34
13	rar	16990	0.34
14	ps	8590	0.17
15	mov	8388	0.17

Table 6: Different file categories specified in queries.

File category	Percentage (%)
Video	30.01%
Audio	24.15%
Compressed file	2.01%
Graphic	1.12%
Document	0.67%
Software	0.36%
Other	41.69%

5. Ways to Reduce P2P Network Traffic

From the results of this study, it has become evident that excessive P2P network traffic is caused mainly by repeat queries. Clearly, the repeat queries are generated partly by the user's inputs and partly by the P2P servant. This excess is achieved by repeating the previous queries periodically if the earlier queries receive no response. Reducing the number of repeat queries could effectively resolve the problem of network traffic. Several proposals have been presented in this paper with suggestions for reduction of repeat queries.

User Generating Repeat Queries

(1) **Precise query**: Earlier location of desirable files by the users lowers the chance of issuing the repeat queries. Technically, this implies the use of file meta data to be used in queries. XML queries could also be another solution to this problem.

P2P Servant Generating Repeat Queries

(1) **Cache Mechanism**: Query messages are propagated from one peer to another peer within the P2P network. A cache mechanism similar to Markatos [5] can provide a useful solution.

(2) **Auto-query Feature**: The ability to disable the auto-query generating feature of P2P servants will reduce some unnecessary and redundant queries to be processed by other P2P servants.

(3) **Web Searching Mechanism**: Precise query means a longer query length, for example "Minority Report Tom Cruise". This may also require query operators, "AND", "OR", and so on. Matching algorithm is another necessary component in this set up. However, application of web searching technologies to a local-based P2P servant may be impractical.

(4) **Query Result**: By moving the "trendy" items to the top of the list (sorted by the released date) and video files (determined by the file extension) the resulting set may be helpful to the P2P users.

6. Conclusion and Future Research

In this paper, searching behavior has been used to address the P2P network traffic problem. The authors studied searching behavior by examining a 7-day long log file. The log file contains all QUERY messages going through a Gnutella-based P2P servant. The results have revealed a P2P phenomenon, identified as a network traffic overflow caused by queries, 97.5% of which are repeat queries. This paper has advocated several ways to reduce repeat queries. The suggested methods are by no means new. However, these methods are justifiable given

the findings of the present study on searching behavior. With this insight into searching behavior, other solutions related to management and technology are possible.

One plausible direction for future research could be utilization of knowledge about the sharable files currently available in the P2P network for further reduction in repeat queries. This can be achieved by studying the QUERYHIT messages in the P2P network.

The findings may also have impacts to improve P2P search. The below is a brief summary of the implications.

Precise query: *Pros - reduced repeated queries* *Cons – fewer results will be obtained*
Cache Mechanism: *Pros – network traffic can be reduced* *Cons – return outdated result, storage space is needed for each peer*
Auto-query Feature: *Pros – reduced repeat query* *Cons – longer time to get desirable result*
Web Searching Mechanism: *Pros – more precise queries* *Cons – getting undesirable results due to misuse of search operators*
Query Result *Pros – help users to locate the desirable results faster* *Cons – the results may not be sorted in the way that the users want*

7. Acknowledgments

This research is supported in part by the Teaching Development Grant (TDG), and the Sino Software Research Institute (SSRI) (Ref. SSRI01/02.BM01). We would like to thank Jack Wong, Elsa Wong, Derek Hau, Alice Chiang, Henry Lin, Ivan Ng, and Jennifer Chan for the implementation of the P2P prototype.

8. References

[1] E. Adar and B. A. Huberman, "Free Riding on Gnutella," *First Monday*, vol. 5(10), 2000.

[2] Anonmyous, Jtella, http://www.kenmccray.com/jtella/, accessed on 1 August 2002.

[3] Clip2.com, The Gnutella Protocol Specification v0.4, http://www9.limewire.com/developer/gnutella_protocol_0.4.pdf, accessed on 1 August 2002.

[4] B. J. Jansen and U. Pooch, "A Review of Web Searching Studies and a Framework for Future Research," *Journal of the American Society of Information Science*, vol. 52(3), pp. 235-246, 2001.

[5] E. P. Markatos, "Tracing a large-scale peer to peer system: an hour in the life of Gnutella," presented at Cluster Computing and the Grid 2nd IEEE/ACM International Symposium CCGRID2002, 2002.

[6] R. Matei, A. Iamnitchi, and P. Foster, "Mapping the Gnutella network," *Internet Computing, IEEE*, vol. 6(1), pp. 50-57, 2002.

[7] S. Saroiu, P. K. Gummadi, and S. D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," presented at Proceedings of the Multimedia Computing and Networking (MMCN) 2002, San Jose, CA, USA, 2002.

[8] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic, "Searching the Web: The Public and Their Queries," *Journal of the American Society of Information Science*, vol. 52(3), pp. 226-234, 2001.

[9] S. Thadani, Meta Information Searches on the Gnutella Network, http://www.limewire.com/index.jsp/metainfo_searches, accessed on 8 August 2002.

[10] J. Vaucher, P. Kropf, G. Babin, and T. Jouve, "Experimenting with Gnutella Communities," presented at Distributed Communities on the Web (DCW 2002), Sydney, Australia, 2002.