Running Head: P2P SEARCHING TRENDS

P2P Searching Trends: 2002 – 2004

Sai Ho KWOK
Department of Information Systems
California State University, Long Beach
College of Business Administration
California State University, Long Beach
Long Beach, CA 90840-8506, USA
jkwok@ust.hk

Abstract

This paper presents the trends of searching queries by users from Peer-to-Peer (P2P) networks over an eighteen-month period from July 2002 to January 2004. Four data sets of search queries collected from Gnutella were studied to describe the searching trends. Major findings include (1) the percentage of duplicate queries ranging from 34% to 68% of total queries; (2) an increase in non-English queries; (3) an approximately half of searching queries specified for video or audio file types; (4) the stop word "the" accounting for one-third of total stop words; (5) the shift of queries from audio to video; and (6) P2P users demanding for timely entertainment and porn materials. Based on the findings, it is worthwhile for P2P developers to consider (1) system design that allows effective searching using multiple languages; and (2) techniques that eliminate stop words for faster searching.

Keywords: Peer-to-Peer; Searching Trends; Queries

1. Introduction

Peer-to-peer (P2P) computing has attracted great interest and attention of the computing industry and gained popularity among computer users and their networked virtual communities. Yang and Garcia-Molina (2002) defined peer-to-peer systems as "distributed systems in which nodes of equal roles and capabilities exchange information and services directly with each other" (p. 1). Anonymity, scalability, fault resilience, decentralization and self-organization are the distinct characteristics of these P2P systems (Milojicic et al., 2002). In Gnutella-based P2P systems, searching is performed using message exchange. However, such message-based searching mechanism results in a large number of flooding query messages over the Gnutella networks. Many P2P researchers begin to address this problem and several solutions have been proposed. Matei, Iamnitchi and Foster (2002) reported the network usage and traffic level of a public Gnutella network and called for smarter routing mechanisms. Besides this study, Tetsuya, Sakai, Kikuma and Kurokawa (2003) suggested that applications of IP multicasting help reduce network traffic.

This paper does not intend to propose any immediate solutions to the problem of Gnutella flooding traffic but aims to report on the searching query trends in the past two years. It is believed that an effective solution to the network traffic problem should integrate human behavior into design procedures / all stages of future development and

implementation. Besides, P2P searching trends have not been explicitly reported in previous studies and relevant data are lacking in the literature. Therefore, there is a need for a more detailed analysis of the searching trends. This study investigates the changes and trends in Gnutella-based searching queries from July 2002 to January 2004. Our research design follows several well-known Web search studies (Jansen, Spink and Saracevic, 2000;Ozmutlu, Ozmutlu and Spink, 2003;Silverstein, Henzinger, Marais and Moricz, 1999;Spink and Ozmultu, 2002;Wolfram, Spink, Jansen and Saracevic, 2001).

2. Background

2.1 Overview of Gnutella Network

In a Gnutella network, each peer is directly connected to one or more peer(s) that are known as neighbors. A peer maintains information repository available for sharing. The shared information is usually about multimedia files. A peer generates a search query and broadcasts the query to its neighbors through a flooding-based broadcast mechanism. Its neighbors check their own shared files in an attempt to locate matched or relevant files, and at the same time, broadcast the incoming query to their connecting neighbors except the peer from which the query is being sent. The searching query

continues to travel along different paths of different peers until the TTL (Time to Live) value in the searching query reaches zero.

2.2 Query Messages of Gnutella Network

There are five types of message or descriptor according to Gnutella Specification 0.4 (Gnutella, 2004) that are used for communicating data between peers on a Gnutella network. They are Ping, Pong, Query, QueryHit and Push messages. A Query message is used for specifying criteria for files while a QueryHit message contains a list of sharable files matching the given criteria returning from neighbors. In this study, we focused on the Query message only because it can indicate the nature of searching queries of P2P users.

2.3 Prior studies on User Interactions with Information Retrieval (IR) systems.

Many studies were conducted to investigate the user interactions with Web search engines and other information retrieval (IR) systems. A growing number of studies have investigated the nature of queries on WWW search engines in the past few years. Jansen, Spink and Saracevic (2000) reported the result of a study on 51,474 queries from 18,113 users of Excites search engines focusing on sessions, queries, and terms. There were 18,098 unique queries among 51,474 queries and 21,862 unique terms

among 113,746 terms. They reported that the sessions were short and most users searched with one query per session only. The queries were short and 44.78% of unique terms appeared once only. Silverstein, Henzinger, Marais and Moricz (1999) reported the result of a study of 154 millions unique queries of Alta Vista search engine. The report shows that the sessions were also short and users mostly typed in short queries. Wolfram et al.Wolfram, Spink, Jansen and Saracevic (2001) reported similar results in another study on 1 million queries of Excite search engine. Ozmutlu, Ozmutlu and Spink (2003) also found that multitasking information seeking and searching was a common behavior, based on the data from AllTheWeb (FAST) search engine. Multitasking sessions often included more than 3 topics per session. They investigated the characteristics of question format of Ask Jeeves search engine (Spink and Ozmultu, 2002) where 30,000 queries were included in the study. The questions were mainly in "where", "what", or "how" format. There are studies on changes and trends of searching queries of WWW search engines. Wolfram, Spink, Jansen and Saracevic (2001) compared the results of the studies of Excite queries from between 1997 to 1999. Findings included that there were fewer terms per query, fewer queries per session and little modification in subsequent queries. Later, Ozmutlu, Ozmutlu and Spink (2003) further compared the results of Excite queries among 1997, 1999, and 2001, specifically for multimedia queries identified. They revealed that queries per

multimedia session had decreased since 1997 and multimedia queries identified were longer than non-multi-media queries.

3. Research Goal

Although there are a number of studies on queries of WWW search engines, studies on queries over P2P networks have seldom been reported and studies on P2P searching trends have never been reported yet. Therefore, the research goal of this study is to investigate the trends in P2P searching, on the basis of searching queries by Gnutella users from July 2002 to January 2004, focusing on the changes in the content of interest, use of language, specific file types and usage of terms. To assess the changes in these searching characteristics and thus trends in P2P searching, we analyzed four weekly data sets, which were captured in July 2002, December 2002, September 2003 and January 2004 respectively. The analysis focused on the characteristics of the searching query, including composition of queries in terms of duplication and use of language, popular file types, queries and terms, etc. By comparing the query characteristics in different sets of data, we are able to assess the changes in searching query and trends in P2P searching. We identify the factors behind the trends, such as copyright infringement and P2P user community. We also discuss

future development of P2P networks and the implications to P2P development, such as reducing P2P network traffic.

4. Research Design

In this section, we describe issues related to data collection and briefly explain several important terms coined in our study.

4.1 Data Collection

Markatos (2002) showed that the overall P2P network characteristics, which are independent of location, can be represented by a randomly-chosen peer in the P2P network and every peer exhibits similar network characteristics. Similarly, we applied these networking characteristics in data collection. Four sets of data were captured in the periods of 9th – 15th July 2002 (labeled Jul2002), 30th December 2002 – 5th January 2003 (labeled Dec2002), 8th – 14th September 2003 (labeled Sep2003) and 6th – 12th January 2004 (labeled Jan2004). Before implementing the data collection, a P2P program was written in JAVA language and implemented in the Gnutella protocol. The P2P program was situated in the campus of the Hong Kong University of Science and Technology and running on a Pentium4 1.6G PC with 10Mps bandwidth. All searching queries passing through the P2P program were logged in the log file. Sample logged

data are presented in Table 1. The log file was then imported into a database management system for further data processing and analysis.

4.2 Definition of Terms

Several terms pertinent to this study are defined as follows:

GUID and Search Criteria

GUID is a 16-byte string uniquely identifying the (message) descriptor on the Gnutella network. Search criteria refer to the criteria input by end-users, in order to search for particular files. Common examples of search criteria are the names of movie and artist. The search criteria and terms are all case-insensitive.

English and Non-English Queries

An English query contains only characters that are subsets of regular expressions including English characters, numeric characters, etc. In contrast, a Non-English query contains Non-English characters such as Simplified Chinese and Korean characters. Any query containing Non-English characters is classified as a Non-English query even the query includes English characters.

Duplicate and Distinct Queries

Classification of duplicate queries is based on the combination of GUID and search criterion. Duplicate queries stand for queries with identical combination of GUID and search criterion. Two queries are classified as distinct queries even if they have identical GUID but distinct search criterion, and vice versa. Queries are classified as duplicate queries in case it appears more than once in the data sets. For example, if a query appears twice out of ten queries, one query will be classified as duplicate queries and the percentage of duplicate queries is 10%. There are duplicate queries because these "identical queries" are routed through different neighboring hosts to the same host by the broadcast-based search mechanism.

Terms

Terms refer to strings of characters bounded by white space.

Stop Words

Stop words are extremely common words that are usually ignored by major search engines such as Google (Google, 2004) at the time of searching, so as to speed up searching. The terms 'the', 'of' and 'in' are typical examples of stop words. In this study, we classified a set of common terms as stop words as listed in Table 2.

5. Comparison of Characteristics of Queries

In this study which spanned 18 months, four weekly sets of searching queries were collected; they were labeled Jul2002, Dec2002, Sep2003 and Jan2004 respectively. Searching queries were then compared and analyzed by the following characteristics including (1) percentage of duplicate queries, (2) percentage of non-English queries, (3) file types, (4) analysis of top 10 queries, (5) analysis of top 20 terms, (6) stop words usage and (7) number of terms per query.

5.1 Duplication of Queries in the four data sets

The number of searching queries in the four sets of data labeled Jul2002, Dec2002, Sep2003 and Jan2004 totaled 4.87 millions, 8.97 millions, 3.05 millions and 3.37 millions respectively.

As depicted in Table 3, the percentage of duplicate queries in all four data sets ranged from 34% to 68% of total queries. Although the percentage of duplicate queries had increased from 44% in July 2002 to 52% in December 2002 and finally to 67% in September 2003, the duplication percentage dropped to 34% in January 2004. This implies that a peer on the Gnutella-like network is likely to receive the same searching message (with the same GUID and search criterion) at a probability of more than 34%. The locality exhibited by the four sets of searching queries in our study is similar to the study by Markatos (2002) who reported that a Gnutella peer receives the same query

message more than once at a probability of approximately 40%.

5.2 Percentage of Non-English Queries

Table 4 shows the percentage of non-English queries on the four data sets. Our analysis of the composition of queries in terms of language and content was based on the distinct queries sets extracted from the raw log files. This ensures a precise assessment of a user's searching behavior, which is free from the interference of locality.

The percentage of non-English queries increased sharply from 0.23% in July 2002 to 12.78% in September of 2003, and then from 12.78% in September of 2003 to in 22.59% in January of 2004. This reveals an increase in the number of queries containing non-English characters in an attempt to locate the files named in languages other than English. One possible reason is that there is an increasing number of shared files named in languages other than English. Another possible reason is that non-English users whose first languages are not English have been on the rise and therefore they constitute a higher portion of P2P communities. In general, P2P users are likely to name and search files with their native languages.

Our study further notes that the queries containing Simplified Chinese characters have appeared on the list of top 10 queries during September 2003 and January 2004.

This may imply that the growth rate of Chinese P2P users (mainly residing in Mainland China) is faster than the growth rate of English-speaking P2P users (mainly residing in the United States). According to a report of CNNIC (CNNIC, 2004), on the one hand, the China Internet population had increased from 45.8 millions in July 2002 to 59.1 millions in January 2003 (at a 29% growth rate), and finally reached 68 millions in July 2003 (at a 15% growth rate). On the other hand, there was almost no growth of the American Internet population during the year 2002, followed by a slight growth during the first two quarters of year 2003 to reach 126 millions Internet population (Rainie and Madden, 2004).

5.3 File Types

When analyzing the file type searched by P2P users, one interesting finding is that there were approximately 40% of searching queries not specifying any file types over the four different periods, as shown in Table 5. One possible explanation is that a significant portion of P2P users would be made up of non-proficient computer users who are not aware of the need to specify desired file types in their queries. These non-professional home users may have constituted a considerable population of P2P communities after P2P file-sharing systems have been widely used since the appearance of Napster.

Another interesting finding is that video and audio file types were mostly clearly specified in the searching queries from users. As most of the audio queries specified "mp3" format as their targeted files, this shows that the "mp3" format has been the most well-known and popular audio file format among P2P user communities in recent years. The portion of searching queries specifying audio files had slightly increased from 24.62% in July 2002 to 26.53% in December 2002, yet followed by gradual decreases to 24.14% in September 2003 and finally 18.64% in January 2004. The significant reduction in audio queries during 2002 could probably result from the RIAA lawsuits against online music file sharers. On 25th June 2003, a day in between the data collection periods of December 2002 and September 2003, the RIAA had announced to sue P2P users who shared music files for copyright infringement. RIAA did take action to file lawsuits against the suspected P2P users on 8th September 2003, a day in between September 2003 and January 2004. A survey conducted by The Pew Internet & American Life Project (Rainie and Madden, 2004) showed that the percentage of American Internet users engaging in sharing music files in P2P network dropped from 35 millions during March, April and May 2003 to 18 millions during November and December 2003. As the report stated, the RIAA lawsuits against online music file sharers appeared to have had a devastating impact on the number of those engaging in Internet P2P music sharing.

In contrast to audio queries, specifying video files has shown an increasing trend throughout all periods. Beginning from 25.09% in July 2002, the percentage of video files finally reached the peak of 35.20% in January 2004. This suggests P2P users have increasing interest towards video files. We also found that searching queries specifying "avi" and "mpg/mpeg" constituted the majority of video queries.

Other file types searched by P2P users include Document, Image and Compressed files. Since queries of these file types accounted for only a small percentage of total queries, about 5 to 7.5%, we did not present these insignificant figures in this study.

5.4 Top 10 Queries

The top 10 queries and their percentages are shown in Table 6. In terms of content, most of the top queries exhibited P2P users' strong interest in entertainment topics over the four different periods. This implies that users' interests remained more or less the same throughout the 1.5 years duration. Usually, entertainment queries could be the name of recently released movie ("Spiderman", "the Lord of the Rings" and "The Last Samurai"), name of artist ("Eminem") and name of anime ("macross flashback").

Another finding that draws our attention is that the porn queries have not been among the top 10 queries since September 2003 although some users had explicitly

searched for porn materials in the data sets captured during 2002. At the initial phrase of our project in mid 2002, three queries explicitly searching for porn materials ("porn", "porn mpg" and "sex") were recorded while two queries looking for erotic materials ("porn" and "sex") were captured at the end of 2002.

Besides entertainment and sexual-related queries, several popular queries that only specified the file type, but with no hint of the content were captured. Generally speaking, those top searching queries specifying file types were all specifically looking for video files ("divx" and "divx avi", "rm" and "dvd").

In terms of the use of language, there were no popular queries in language other than English in all the data sets collected during the six-month period in the year 2002. However, there were three top queries in September 2003 and two top queries in January 2004 containing Simplified Chinese characters.

5.5 Top 20 Terms

In this section, we compare and analyze the most frequently used terms in the distinct queries in the four different periods. Our analysis of the top 20 terms excluded the stop words listed in Table 2.

With reference to Table 7, it is observed that the extensions of audio and video files that were frequently specified in searching queries account for the top five terms for all periods. These terms include "mp3", "avi" and "mpg".

Our findings revealed a total of 12 terms commonly appeared among the top 20 terms in the four different periods. They are "1", "2", "-", "avi", "i", "love", "mp3", "mpg", "my", "porn", "sex" and "xxx". Among them, two terms are probably related to audio files; "-" is usually the separator between an artist and his/her song; "mp3" is the most common audio file format; and "1" as well as "2" probably are the track numbers of an album. Moreover, the terms "avi" and "mpg" are directly linked to video files. We also noted three terms which are related to porn materials, "porn", "sex" and "xxx". The remaining terms are general terms "i", "love" and "my" that when used alone is unlikely to effectively locate the desired files. Based on the most frequent terms that are common in the four different periods, P2P users could be said to focus their interest and attention narrowly and continuously on video files, audio files and porn materials over the 1.5 years duration.

5.6 Stop Words

Our study found that the stop words generally accounted for 6-8% of the total terms used, with a distribution of 6.96% in July 2002, 7.41% in December 2002, 6.41% in September 2003 and 7.01% in January 2004.

As depicted in Figure 1, the four data sets captured at different periods of time all exhibited a similar pattern on the distribution of stop words. The stop word "the" is the most frequently used stop word in searching queries and accounts for approximately one-third of the total stop words. The second most frequent stop word is "of" that contributes to about 15% of total stop words. The stop words "a", "and", "in" and "to" are other widely used stop words. The percentages of most frequent stop words are listed in Table 8.

5.7 Number of Terms per Query

Different data sets of searching queries have exhibited significant differences on the number of terms per query. The average number of terms per query is 3.96 in July 2002 and increased to 6.47 in December 2002. Yet the mean number of terms per query had fallen to 5.63 in September 2003 and finally dropped to 4.74 in January 2004. The distribution of number of terms per query demonstrated a fluctuating trend throughout the 1.5 years duration. As illustrated in Figure 2, the distributions of number of terms per query were left-skewed in July 2002, September 2003 and January 2004

such that a significant portion of queries contained either two terms or a single term. However, the number of terms per query in December 2002 ranged from one term to eight terms.

Our analysis of the top 10 queries clearly shows that users were consistently interested in timely entertainment contents such as recently released movies. The varying numbers of terms per query recorded during the different periods were probably related to the name of timely contents. For example, the recently released movies "Spiderman" in July 2002 and "The Lord of the Rings: The Two Towers" in December 2002.

6. Discussion

From our study, we discovered that more than one-third of the searching queries are duplicate queries. Such problem of locality reduces the efficiency of network utilization and places unnecessary network traffic on the infrastructure of Internet Service Providers (ISPs). The impact on the ISP will be more serious as P2P users exhibit increasing interests towards video files that are generally of large file size. Such heavy P2P network traffic brought by duplicate queries could be reduced by employing a caching mechanism, which temporarily stores and reuses the query results of popular queries, such as audio and video queries. Without the caching mechanism, a peer who

receiving a query will forward it to all its neighboring peers and pass the query results back to the requesting peer. With the caching mechanism, a peer can store the query results of the popular queries temporarily. When the peer receives similar queries, the peer does not have to forward the query to other peers but responds to the requesting peer with the query results residing in cache. The peer is required to update its cache occasionally in order to provide accurate query results. Such a caching mechanism not only saves the network bandwidth for distributing duplicate queries, but also provides a faster response to popular queries. To achieve this, an intelligent agent can be implemented in P2P client programs.

Our findings also show that large files such as songs and movies are demanding in P2P file sharing. In the current P2P network design, a simple byte-by-byte downloading mechanism is used to retrieve all kinds of files, regardless of their sizes and types. With the rapid growth of shared files, it could be efficient to introduce different operating platforms for different types of file, and apply different downloading mechanisms to them. For example, P2P video streaming could be a better way of distributing video files than downloading video files byte by byte. This could better utilize the network bandwidth. In addition, separating files based on their types could facilitate efficient file searching because P2P users could locate their desired files

more easily and precisely when they perform searches in the corresponding file platform.

From the above findings, we believe that capturing the searching trends enables us to understand the user behavior better, and further apply it to the design of P2P systems and integrate it into ISP regulations for effective bandwidth utilization.

Future potential research topics include the changes in user search behavior over the day, the reformulation of queries and the changes in search behavior when the users perform multitasking search.

7. Conclusion

This paper reported the searching query trends of a Gnutella-based P2P network from July 2002 to January 2004. The trends cover (1) the percentages of duplicate queries; (2) the percentages of non-English queries; (3) the percentages of queries for video and audio; and (4) the percentages of stop words. Based on the findings, we believe that it is worthwhile for P2P developers to consider (1) system design that supports effective searching features using multiple languages; and (2) techniques that eliminate stop words for faster searching. In addition, our findings could be beneficial for ISP to save bandwidth for other cost-effective activities by regulating P2P activities.

8. Acknowledgement

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

CNNIC (January 2004). 13th Statistical Survey on the Internet Development in China. http://www.cnnic.net.cn/download/manual/en-reports/13.pdf.

Gnutella. Gnutella Specification 0.4. http://www.stanford.edu/class/cs244b/gnutella_protocol_0.4.pdf.

Google. Google. http://www.google.com.

B. J. Jansen, A. Spink, & T. Saracevic (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. Information Processing & Management, 36 (2), 207-227.

E. P. Markatos (2002). Tracing a large-scale peer to peer system: an hour in the life of Gnutella. In IEEE/ACM International Symposium on Cluster Computing and the Grid (pp. 65-74).

R. Matei, A. Iamnitchi, & P. Foster (2002). Mapping the Gnutella network. IEEE Internet Computing, 6 (1), 50-57.

D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, & Z. Xu, "Peer-to-Peer Computing," HP Laboratories Palo Alto HPL-2002-57, 2002.

S. Ozmutlu, H. C. Ozmutlu, & A. Spink (2003). A study of multitasking web search. In International Conference on Information Technology: Coding and Computing

(ITCC2003) (pp. 145-148).

L. Rainie & M. Madden (4 January 2004). The impact of recording industry suits against music file swappers. http://www.pewinternet.org/reports/pdfs/PIP_File_Swapping_Memo_0104.pdf

C. Silverstein, M. Henzinger, H. Marais, & M. Moricz (1999). Analysis of a Very Large Web Search Engine Log. ACM SIGIR Forum, 33 (3).

A. Spink & H. C. Ozmultu (2002). Characteristics of question format Web queries: an exploratory study. Information Processing & Management, 38 (4), 453-471.

O. i. Tetsuya, K. Sakai, K. Kikuma, & A. Kurokawa (2003). Study of the relationship between peer-to-peer systems and IP multicasting. IEEE Communications Magazine, 41 (1), 80-84.

D. Wolfram, A. Spink, B. J. Jansen, & T. Saracevic (2001). Vox populi: the public searching of the Web. Journal of the American Society for Information Science & Technology, 52 (12), 1073-1074.

B. Yang & H. Garcia-Molina (2002). Improving search in peer-to-peer networks. In 22nd International Conference on Distributed Computing Systems (pp. 5-14).

Table 1

The first entry is extracted from the data set July2002 and the second entry is taken from Sep2003 data set; The "Date Time" indicates the capture time; the "Hops" refers to the number of hosts the searching query passed through; the "GUID" refers to the unique identity of a message.

Date Time	**TTL**	**Hops**	**GUID**	**Search Criteria**
12/9/2003 7:03	1	6	[70]-[28]-[d]-[7]-[5c]-[1c]-[19]-[4f]-[99]-[25]-[94]-[a3]-[21]-[19]-[eb]-[82]	down with love
2/7/2002 3:15	1	6	[15]-[38]-[ba]-[5b]-[7a]-[2d]-[5d]-[4a]-[96]-[2e]-[e0]-[d2]-[12]-[b8]-[94]-[40]	cartoon mp3

Table 2

The list of stop words

&	a	an	and	are	at	be	by	for	from	if
in	into	is	of	on	the	through	to	up	which	with

Table 3

The percentages of distinct and duplicate queries

	July 2002	December 2002	September 2003	January 2004
Total Queries	4,876,773	8,970,889	3,053,418	3,373,171
Percentage of Distinct Queries	55.84%	47.51%	32.38%	65.62%
Percentage of Duplicate Queries	44.16%	52.49%	67.62%	34.38%

Table 4

The percentage of Non-English queries

	July 2002	December 2002	September 2003	January 2004
Percentage of non-English Queries	0.23%	0.14%	12.78%	22.59%

Table 5

Distribution of queries specifying file types; percentage of top three audio file types and top three video file types are also listed.

	July 2002	December 2002	September 2003	January 2004
Audio Types	24.62%	26.53%	24.14%	18.64%
Audio Type 1 (%)	mp3 (24.38%)	mp3 (25.86%)	mp3 (23.62%)	mp3 (18.13%)
Audio Type 2 (%)	wav (0.17%)	mid** (0.24%)	wma (0.30%)	wav (0.20%)
Audio Type 3 (%)	wma (0.05%)	wav (0.21%)	wav (0.20%)	wma (0.18%)
Video Types	25.09%	27.61%	29.87%	35.20%
Video Type 1 (%)	mpg*(12.31%)	avi (11.33%)	mpg(10.13%)	mpg(12.00%)
Video Type 2 (%)	avi (11.00%)	mpg(11.31%)	avi (9.52%)	avi (9.74%)
Video Type 3 (%)	divx (1.31%)	rm (2.90%)	rm (6.09%)	rm (7.42%)
Others* **	5.06%	7.48%	6.91%	6.62%
Not specified	45.23%	38.38%	39.08%	39.54%
Total	100%	100%	100%	100%

Remarks: * "mpg" stands for "mpg/mpeg"; ** "mid" stands for "mid/midi"; *** refers to other unpopular file types including Document, Image and Compressed files.

Table 6

Top 10 queries from four data sets

Rank	July 2002	December 2002	September 2003	January 2004
1	divx	macross flashback 2012 mpg	影视 rm	LINUX
2	porn	macross flashback 2012 mpeg	Justin Timberlake-So Cool-Justified-18	MOVIE QUOTE
3	eminem	macross flashback	Justin Timberlake_What's A Guy Gotta Do	WALLPAPER
4	divx avi	divx	Justified - You Should Be Dancing	SPEECH
5	SPIDERMAN	porn	Justified - Corners of Your Mouth	the last samurai
6	nelly	eminem	battlefield	PDF
7	Minority report	two towers	级片 rm	The Lord of the Rings
8	Teen	Mussorgsky-Ravel - Pictures at an Exhibition The Great Gate at Kiev	长片 rm	自拍
9	porn mpg	the lord of the rings the two towers 2002 full movie 5 640x352 avi	adobe after effects 6 pro iso multilanguage scotch bin	台湾
10	sex	sex	rm	dvd

Table 7

Top 20 terms in four data sets (excluding the stop words listed in Table 2)

	July 2002		December 2002		September 2003		January 2004	
Rank	Term	Percentage	Term	Percentage	Term	Percentage	Term	Percentage
1	mp3	6.53%	-	7.08%	-	3.93%	-	4.37%
2	avi	2.92%	mp3	1.44%	mp3	2.92%	mp3	2.49%
3	mpg	2.26%	avi	0.96%	avi	1.37%	avi	1.30%
4	zip	0.60%	mpg	0.58%	mpg	1.26%	mpg	1.29%
5	2	0.55%	sex	0.46%	rm	0.93%	wmv	0.78%
6	1	0.52%	you	0.41%	2	0.61%	rm	0.68%
7	you	0.46%	porn	0.37%	1	0.50%	sex	0.64%
8	I	0.45%	i	0.37%	wmv	0.45%	xxx	0.56%
9	mpeg	0.45%	1	0.37%	You	0.41%	Porn	0.53%
10	xxx	0.38%	xxx	0.37%	sex	0.39%	1	0.51%
11	-	0.36%	2	0.35%	zip	0.36%	2	0.50%
12	sex	0.36%	teen	0.30%	i	0.35%	anal	0.41%
13	jpg	0.34%	my	0.27%	xxx	0.34%	Sepultura	0.38%
14	porn	0.34%	me	0.27%	porn	0.34%	teen	0.35%
15	me	0.34%	love	0.25%	me	0.28%	fuck	0.33%
16	love	0.34%	girl	0.22%	love	0.26%	MOVIE	0.32%
17	divx	0.33%	anal	0.21%	my	0.26%	i	0.31%
18	star	0.31%	fuck	0.19%	mpeg	0.23%	lord	0.31%
19	black	0.28%	young	0.19%	teen	0.23%	girl	0.31%
20	my	0.28%	girls	0.18%	s	0.22%	my	0.26%

Table 8

The top 6 stop words and their percentage

	July 2002	**December 2002**	**September 2003**	**January 2004**
the	32.25%	31.12%	31.74%	35.17%
of	15.58%	13.56%	15.74%	16.61%
in	8.70%	7.94%	7.45%	6.29%
and	8.28%	8.14%	6.65%	6.79%
a	7.78%	7.42%	7.43%	6.98%
to	5.50%	5.13%	6.42%	4.17%

Figure captions

Figure 1. The distribution of stop words.

Figure 2. The distribution of number of terms per query.

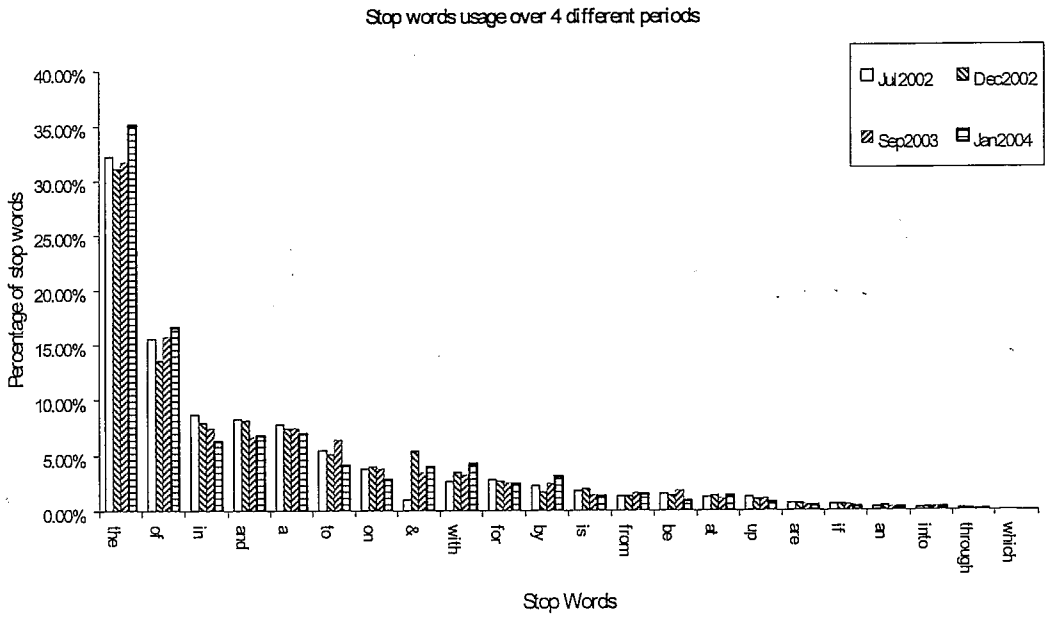

Figure 1. The distribution of stop words.

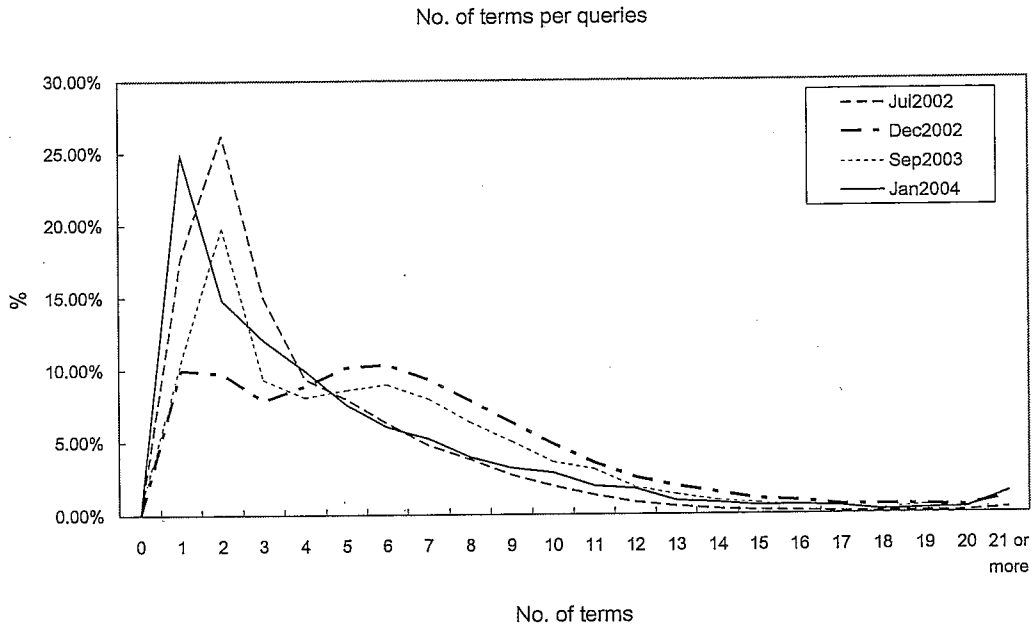

Figure 2. The distribution of number of terms per query.