Changes in queries in Gnutella peer-to-peer networks

Christopher C. Yang

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

James S.H. Kwok

Department of Information Systems, College of Business Administration, California State University, Long Beach, USA

Received 2 June 2004
Revised 7 November 2004

Abstract.

Peer-to-peer (P2P) networks have been drawing significant attention in the area of information management and sharing recently. Unlike the World Wide Web (WWW), every peer in a P2P network is both client and server. It hosts information for sharing and submits queries to search for information from all of the other peers at the same time. There are a large number of studies related to the information behavior of WWW search engines in the US and Europe. Several issues have been investigated in these studies including number of unique and repeat queries, length of queries, terms being used in queries, successive searching, multitasking searching, multiple searching sessions, changes of topics, multimedia searching, etc. The number of similar studies in P2P networks is significantly less. We have previously studied the information behavior of Gnutella P2P network based on data collected in 2002. In this paper, we present changes in Gnutella queries, based on the previously collected data and new data collected in 2003. Similar metrics to those used by Spink, Wolfram, Jansen, and Saracevic and in our previous study are used. We found that the number of non-English queries has increased. The number of repeat queries in P2P has decreased but is still more than that for WWW search engines. The length of queries has been increased by about 40%. The topics of interest have shifted from entertainment and sexuality to computers and entertainment. In general, the information behavior of P2P users has been changing as the P2P technology becomes more mature and the P2P users become more familiar with the technology. As the P2P technology begins to be adopted in e-Business, we foresee that there will be more changes in the future.

Keywords: peer-to-peer networks; World Wide Web; information behaviour

1. Introduction

Peer-to-peer (P2P) networking technology can be described as a mechanism to facilitate file sharing without any centralized coordination. The architecture of the P2P network is classified as pure and hybrid [1]. In the pure P2P network, nodes also known as servants, simultaneously act as both servers and clients. In the hybrid P2P network, ultrapeers, which possess better computational resources, act as a central entity to provide services to leave nodes that possess less computational resources. Gnutella is an example of a pure P2P network while KaZaa and eDonkey are examples of hybrid P2P networks. In this study, we focus on the pure P2P network, Gnutella network, to understand the information behavior of P2P users who may be interested in any types of information, but not any special group of users.

There have been many studies on the information behavior of users of World Wide Web (WWW) search

Correspondence to: Christopher C. Yang, Associate Professor, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: yang@se.cuhk.edu.hk

engines. However, the number of studies on the information behavior of P2P users is significantly lower. Changes in searching behavior and information needs on WWW were also reported by Spink, Jansen, Wolfram, and Saracevic [2]. It was found that there was a shift in the search topics but only little change in user search behavior. In this paper, we are interested in investigating the changes of information behavior of P2P users between 2002 and 2003 based on two sets of data collected from the Gnutella network. The analysis of the set of data collected in 2002 has been presented previously [3]. We are also interested to identify any difference between the information behavior of P2P users and WWW search engine users based on studies of Excite search engine [2, 4–6].

Human information behavior has been studied in several disciplines. Each discipline has its own focus, including personality in psychology, consumer behavior, health communication studies, information requirements in information systems design, and organizational decision making [7]. In general, researchers are interested in how people seek and use information, the channels to access information, and the factors that encourage or discourage information use. In this paper, we focus on the change of information behavior on Gnutella P2P network and the difference in information behavior in two Internet channels, P2P and WWW.

1.1. Searching mechanisms in peer-to-peer networks and the World Wide Web

P2P and WWW have significantly different searching mechanisms. The traditional WWW search engines rely on a centralized index server while the searching on a decentralized and unstructured P2P network (for the case of Gnutella) rely on forwarding query requests through peers. To better understand the results of our study, we elaborate the difference in the searching mechanisms between P2P networks and WWW search engines in this section.

Searching in the decentralized and unstructured P2P networks, such as Gnutella, relies on the message passing among peers on the dynamic networks to locate the peers that have the requested files [8, 9]. A client in Gnutella is a peer that submits queries. A server is a peer that provides content. A router is a peer that transmits queries and responses if it does not have the file requested. ‘Ping’ messages are sent out by peers to identify other peers on the dynamic networks and a ‘Pong’ message is received from an identified peer. Each peer may only identify a few peers in the neighborhood among all the peers on the networks. However, when queries are submitted, the identified peers may forward the queries to other peers that they have identified. When a peer is searching for a file, it sends out a ‘Query’ message containing some filtering criteria. If the requested file is identified, the peer will respond by a ‘Query hit’ message containing the list of files matching the filtering criteria and the IP address of the peer who is the content provider. Gnutella uses time-to-live (TTL) to control the number of hops that a query can be propagated. If the requested files are not identified after a certain TTL, the query will be terminated. Gnutella adopts ‘owner replication’, that means the requested files will be replicated at the requesting peer when the search is successful. Figure 1 illustrates the searching mechanism of Gnutella P2P networks.

The traditional commercial WWW spiders, such as Excite and Google, are composed of a fetching robot, index server, and search engine [10–13]. The fetching robot gathers homepages from the WWW through the links available on the Web page and submission of Web pages. The Web sites are revisited by the fetching robots at some interval of time. The indexing server extracts keywords and phrases from the documents and indexes them with their universal resource locations (URLs). For each word and phrase, the indexing database contains the URL address of the Web page and the location of the word or phrase in the Web page. The search engine allows users to submit queries by keyword and returns a list of URLs according to their order of relevance to the keywords. There are several popular ranking strategies. Most ranking strategies are based on word frequency, Web site popularity, location of words, connectivity of Web pages, and date of Web page being updated. Google adopts the PageRank

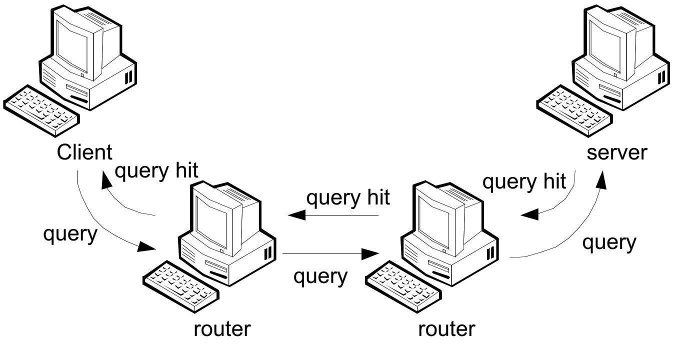


Fig. 1. Searching mechanism of P2P networks, Gnutella.

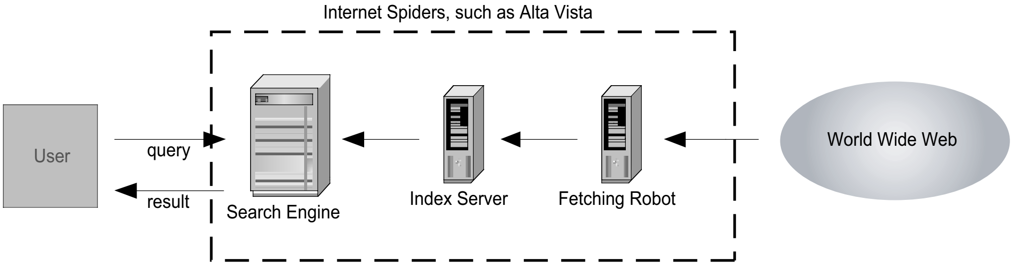


Fig. 2. Searching mechanism of Internet Spiders.

algorithm that is developed based on link analysis. Figure 2 illustrates the searching mechanism of Internet spiders.

As illustrated, the P2P networks searching does not rely on a centralized database to provide the address of the relevant files but relies on its peers to provide the matching files. The results of queries sent to Web search engines depend on the performance of a particular search engine (the power of indexing and fetching and their ranking policies). However, the results of queries sent to P2P networks depend on the protocol of a particular P2P network. They also depend on the peers appearing in the neighborhood and the files being shared by these peers.

1.2. User interfaces of peer-to-peer networks and World Wide Web search engines

Not only the searching mechanisms are different between P2P networks and WWW search engines; their user interfaces for searching, presentation of results and file downloading are also different. Figure 3(a) shows the user interface for searching in the Gnutella network. Users can submit a keyword or name of file in the searching text field and specify the type of files that they are looking for. Other advanced searching features, such as name of artist, album, and genre can also be specified. In Figure 3(a), the searching results are presented in the right frame as a directory tree. For each item, the title of the file, the name of the artist, the file size, the user who provides the file, etc. are presented. Figure 3(b) shows the user interface for downloading selected files. The lower frame of the interface allows users to upload files to share with other users in the Gnutella network.

Figure 4 presents the user interface of Excite search engine for WWW. Users may submit keywords and the results are presented with the title of the Web page, the first few lines of the Web page, and the URL. Ranking of the extracted Web pages is also provided. By clicking on the title of the selected Web page, the Web page will be downloaded and presented on the same browser.

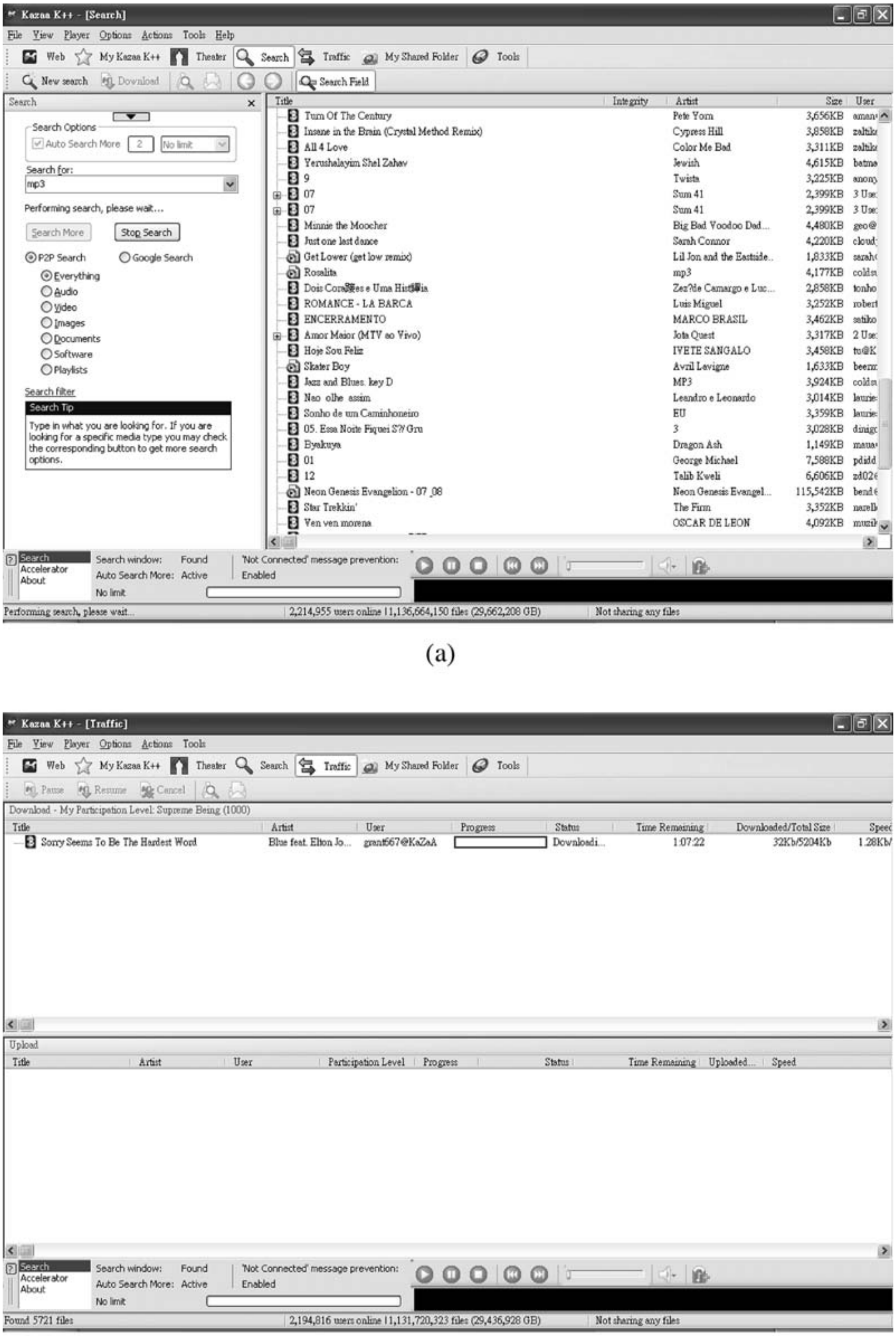
Several major differences between the user interfaces of Gnutella network and Excite search engine can be identified:

- More detailed searching features can be specified in the Gnutella network, for example, the types of files, the names of artists, the titles of videos or songs, etc.
- Results are presented in a different format:
  Directory tree in Gnutella network.
  Ranked list in Excite search engine.
- Results in Excite search engine are hypermedia Web pages with embedded images, videos, and sound clips. Results in Gnutella network are in a specific file format but not hypermedia. The files are audio, video, images, documents, or software.
- The selected result can be downloaded by clicking a link and presented on the same browser for Excite search engine. However, another window for downloading the selected file is provided for Gnutella network.

2. Research design

2.1. Data collection

The data collected for this study was obtained from 1 to 7 September 2003 (seven days) with a P2P servant situated in Hong Kong. As proved by Markatos [14], the characteristic of a P2P network is independent of the location of the nodes. The P2P servant utilized in this study was a Java program based on the JAVA API following the Gnutella protocol [15] and ran on a P4 1.2G PC with 100 Mps bandwidth. The data collected from the query messages was saved in a log file through the Java program and analyzed by a database management system. We label this study 'Gnutella 2003'. A


Kazaa K++ - [Search]
File View Player Options Actions Tools Help
Web My Kazaa K++ Theater Search Traffic My Shared Folder Tools
New search Download Search Field
Search
Search More Stop Search
Search filter
Search Accelerator About

Search Options

☑ Auto Search More 2 No limit

Search for: mp3

Performing search, please wait…

● P2P Search ○ Google Search

● Everything
○ Audio
○ Video
○ Images
○ Documents
○ Software
○ Playlists

Search Tip

Type in what you are looking for. If you are looking for a specific media type you may check the corresponding button to get more search options.

Title	Integrity	Artist	Size	User
Turn Of The Century		Pete Yorn	3,656KB	aman
Insane in the Brain (Crystal Method Remix)		Cypress Hill	3,858KB	zaltik
All 4 Love		Color Me Bad	3,311KB	zaltik
Yerushalayim Shel Zahav		Jewish	4,615KB	batma
9		Twista	3,225KB	anony
07		Sum 41	2,399KB	3 Use
07		Sum 41	2,399KB	3 Use
Minnie the Moocher		Big Bad Voodoo Dad…	4,480KB	geo@
Just one last dance		Sarah Connor	4,220KB	cloud
Get Lower (get low remix)		Lil Jon and the Eastside..	1,833KB	sarah
Rosalita		mp3	4,177KB	coldsu
Dois Cora[illegible]es e Uma Hist[illegible]ia		Zez?de Camargo e Luc…	2,858KB	tonho
ROMANCE - LA BARCA		Luis Miguel	3,252KB	robert
ENCERRAMENTO		MARCO BRASIL	3,462KB	satiko
Amor Maior (MTV ao Vivo)		Jota Quest	3,317KB	2 Use
Hoje Sou Feliz		IVETE SANGALO	3,458KB	tu@K
Skater Boy		Avril Lavigne	1,633KB	beem
Jazz and Blues. key D		MP3	3,924KB	coldsu
Nao olhe assim		Leandro e Leonardo	3,014KB	laurie
Sonho de um Caminhoneiro		EU	3,359KB	laurie
05. Essa Noite Fiquei S?/ Gru		3	3,028KB	dinigo
Byakuya		Dragon Ash	1,149KB	mauas
01		George Michael	7,588KB	pdidd
12		Talib Kweli	6,606KB	zd02
Neon Genesis Evangelion - 07 ,08		Neon Genesis Evangel…	115,542KB	bend
Star Trekkin'		The Firm	3,352KB	narell
Ven ven morena		OSCAR DE LEON	4,092KB	muzik

Search window: Found
Auto Search More: Active
No limit

'Not Connected' message prevention: Enabled

Performing search, please wait… | 2,214,955 users online 11,136,664,150 files (29,662,208 GB) | Not sharing any files

(a)

Kazaa K++ - [Traffic]
File View Player Options Actions Tools
Web My Kazaa K++ Theater Search Traffic My Shared Folder Tools
Pause Resume Cancel
Search Accelerator About

Download - My Participation Level: Supreme Being (1000)

Title	Artist	User	Progress	Status	Time Remaining	Downloaded/Total Size	Speed
Sorry Seems To Be The Hardest Word	Blue feat. Elton Jo…	grant667@KaZaA		Downloadi…	1:07:22	32Kb/5204Kb	1.28Kb/

Upload

Title	Artist	User	Participation Level	Progress	Status	Time Remaining	Uploaded…	Speed

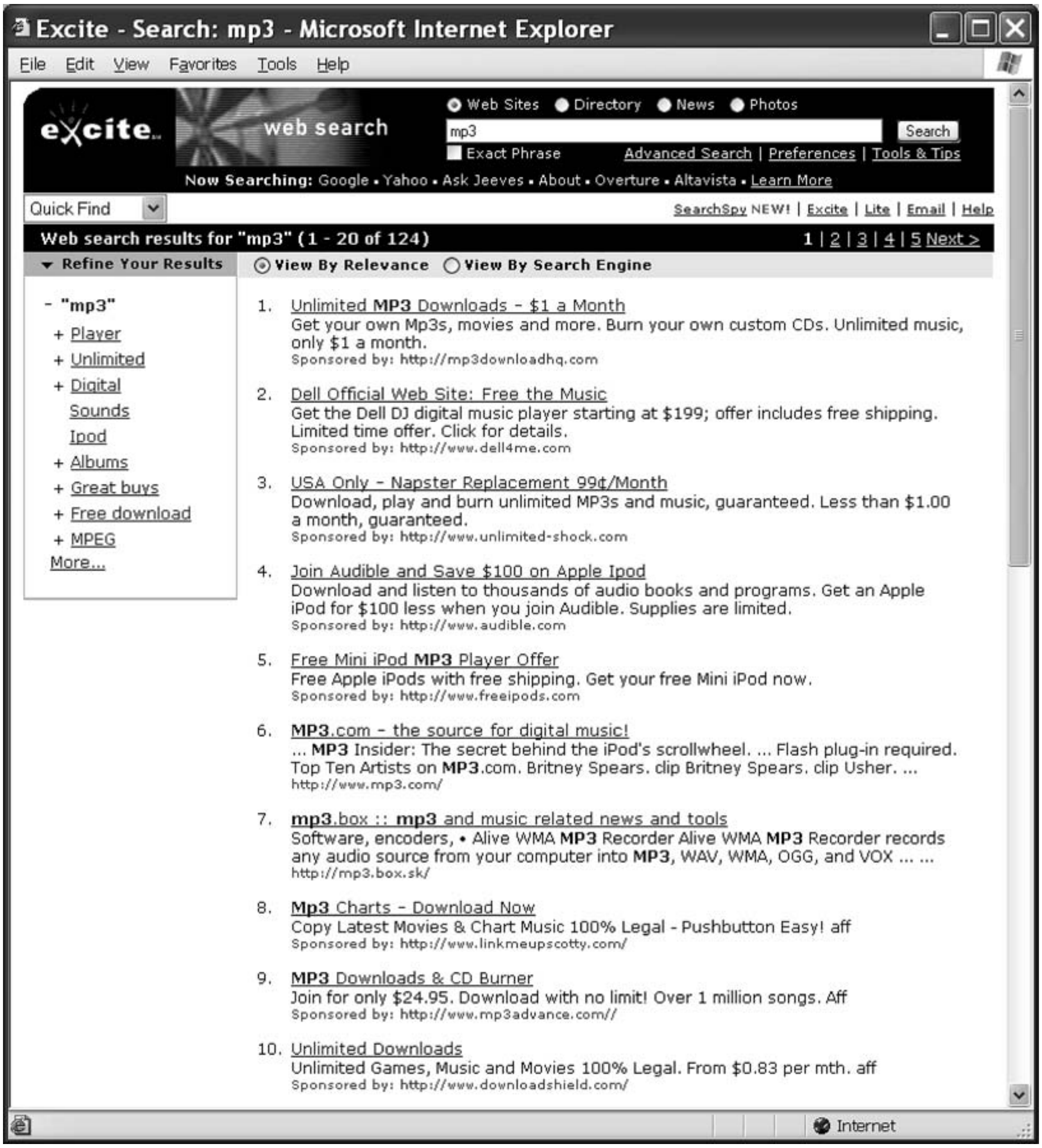
Search window: Found
Auto Search More: Active
No limit

'Not Connected' message prevention: Enabled

Found 5721 files | 2,194,816 users online 11,131,720,323 files (29,436,928 GB) | Not sharing any files

(b)

Fig. 3. User interfaces of Gnutella network, (a) searching and presentation of results, (b) downloading and uploading files.


Excite - Search: mp3 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
excite
web search
● Web Sites ● Directory ● News ● Photos
Search
Advanced Search | Preferences | Tools & Tips
Quick Find
SearchSpy NEW! | Excite | Lite | Email | Help
Internet

mp3

□ Exact Phrase

Now Searching: Google • Yahoo • Ask Jeeves • About • Overture • Altavista • Learn More

Web search results for "mp3" (1 - 20 of 124) 1 | 2 | 3 | 4 | 5 Next >

▼ Refine Your Results

- "mp3"
  - + Player
  - + Unlimited
  - + Digital
  - Sounds
  - Ipod
  - + Albums
  - + Great buys
  - + Free download
  - + MPEG
  - More...

⦿ View By Relevance ○ View By Search Engine

1. Unlimited **MP3** Downloads – $1 a Month
Get your own Mp3s, movies and more. Burn your own custom CDs. Unlimited music, only $1 a month.
Sponsored by: http://mp3downloadhq.com

2. Dell Official Web Site: Free the Music
Get the Dell DJ digital music player starting at $199; offer includes free shipping. Limited time offer. Click for details.
Sponsored by: http://www.dell4me.com

3. USA Only – Napster Replacement 99¢/Month
Download, play and burn unlimited MP3s and music, guaranteed. Less than $1.00 a month, guaranteed.
Sponsored by: http://www.unlimited-shock.com

4. Join Audible and Save $100 on Apple Ipod
Download and listen to thousands of audio books and programs. Get an Apple iPod for $100 less when you join Audible. Supplies are limited.
Sponsored by: http://www.audible.com

5. Free Mini iPod **MP3** Player Offer
Free Apple iPods with free shipping. Get your free Mini iPod now.
Sponsored by: http://www.freeipods.com

6. **MP3**.com – the source for digital music!
... **MP3** Insider: The secret behind the iPod's scrollwheel. ... Flash plug-in required. Top Ten Artists on **MP3**.com. Britney Spears. clip Britney Spears. clip Usher. ...
http://www.mp3.com/

7. **mp3**.box :: **mp3** and music related news and tools
Software, encoders, • Alive WMA **MP3** Recorder Alive WMA **MP3** Recorder records any audio source from your computer into **MP3**, WAV, WMA, OGG, and VOX ... ...
http://mp3.box.sk/

8. **Mp3** Charts – Download Now
Copy Latest Movies & Chart Music 100% Legal - Pushbutton Easy! aff
Sponsored by: http://www.linkmeupscotty.com/

9. **MP3** Downloads & CD Burner
Join for only $24.95. Download with no limit! Over 1 million songs. Aff
Sponsored by: http://www.mp3advance.com//

10. Unlimited Downloads
Unlimited Games, Music and Movies 100% Legal. From $0.83 per mth. aff
Sponsored by: http://www.downloadshield.com/

Fig. 4. User interfaces of Excite search engine – searching and presentation of results.

total of 3,721,024 query messages were collected in Gnutella 2003 over seven days. A query message includes data about date, time, speed, TTL, number of hops, host and search criteria. The search criteria contain information about the queries.

In our previous study [3], the same setup was adopted and data was collected from 17 to 22 July 2002 (7 days). 5,052, 752 query messages were collected. We label it 'Gnutella 2002'.

In this query study, demographic data is not available in the query messages we collect. Indeed, anonymity is a key feature in the Gnutella protocol [8, 16]. In any collection of query messages, no demographic information will be released. Demographic information will only be obtained through the central administrator as the memberships of the P2P network but will not be available as the distribution of peers connected to the P2P network in a particular period of time. The purpose of the anonymity is to encourage information sharing among users. As a result, we are not able to make any comparison of the demographic data between P2P and WWW search engine users.

2.2. *Metrics*

The metrics developed by Spink, Wolfram Jansen, and Saracevic [5] with modifications for P2P network and used in our previous study [3] will be used in this study. Definitions are summarized as follows:

- Terms – unbroken strings of alphanumeric characters including words, abbreviations, numbers and logical operators (for English only). For other languages such as Chinese or Korean, the corresponding coding will be utilized to identify terms. For example, Big5 or GB will be used for Chinese.
- Queries – sets of one or more terms.
- Unique queries – all differing queries entered by one user in a session.
- Repeat queries – all multiple occurrences of the same query by one user or submitted by the system automatically to update the list of results periodically. Users may request the system to submit a query again periodically to update the results since all the peers dynamically connect to or disconnect from the network. Therefore, a query result can change after a period of time.
- Zero term queries – queries without any terms.
- XML queries – queries containing XML substrings to specify the query metadata. For example, the names of artists and the titles of songs or videos can be specified as XML queries.
- English queries – queries containing only English words, numbers and symbols.
- Non-English queries – queries containing non-English words, such as Chinese or Korean characters.

3. Results

3.1. *Changes in Gnutella queries in terms of types*

Table 1 presents the occurrence and distribution of different types of queries in Gnutella 2002 and Gnutella 2003. In Gnutella 2003, there are 80.57% repeat queries and 0.01% zero term queries, which have decreased from 86.56% and 0.05% respectively in Gnutella 2002. On the other hand, there are 19.42% unique queries, which is an increase from 13.39% in Gnutella 2002. Based on the difference between Gnutella 2003 and Gnutella 2002, we can see the trends of decreasing repeat queries and increasing unique queries.

Using similar metrics, studies of queries in Excite search engine for World Wide Web have also been conducted by Jansen et al. [17, 18], Wolfram et al.[6], and Spink et al. [2]. Jansen et al. conducted their study on 9 March 1997 with 51,473 queries collected. We label this study 'Excite 1997'. Wolfram et al. and Spink et al. conducted their study on 20 December of 1999 with 1,025,910 queries collected. We label it 'Excite 1999'. Based on the Excite 1997 and Excite 1999 studies as shown in Table 2, we find that the percentages of unique and repeat queries have slightly increased and decreased respectively, by approximately 3%.

Similar to the Excite studies, there are changes in the percentage of repeat queries and unique queries but the changes are relatively greater. The percentages of unique and repeat queries have increased and decreased, respectively, by approximately 6%. The percentage of repeat queries in Gnutella 2003 is still double that in Excite 1999. As discussed in our earlier study [3], the connection of peers is dynamic. Therefore, users may resubmit queries to update or extend the list of search results. However, the pattern of repeating queries has decreased after one year. It is possible that the users have learned that submitting queries repeatedly does not help much after all. Refining queries or submitting new queries are more helpful. Besides, the P2P technology has become more mature. P2P hosts with XML queries enabled are more popular. The searching performance is more stable. All of these factors encourage the unique queries.

In Table 1, we find that there is a substantial increase in non-English queries from 0.59% in Gnutella 2002 to 20.85% in Gnutella 2003. XML queries also increase, from 0.77% in Gnutella 2002 to 7.23% in Gnutella 2003, but not as significantly as non-English queries. On the other hand, there is a substantial decrease in English queries from 98.64% in Gnutella 2002 to 71.92% in Gnutella 2003. Similar changes in the percentages of English and non-English queries were also reported in the Excite studies [2, 17]. It was reported that non-English or unknown queries increased from 4.1% in 1997 to 6.8% in 1999 and 11.3% in 2001. It is clear that there are more non-English users and non-English information in both WWW and P2P in recent years [19, 20]. Therefore, it is natural that there is a substantial increase in non-English queries. It is worth noting that the increase in non-English queries in P2P is greater than that in WWW.

It is found that the majority of non-English queries are Chinese queries, which are expected to originate from users in China. China has experienced a significant growth in Internet use. According to the *Semiannual Survey Report on the Development of China's Internet* [21] in July 2002 and July 2003, the total number of computer hosts in China was 16.13 million

Table 1
Occurrences and distribution of different types of queries in Gnutella Network collected in 2002 and 2003

	Gnutella 2002 [3]		Gnutella 2003	
	Occurrences	Percentage %	Occurrences	Percentage %
Total Number of Queries	5,052,754	100.00%	3,721,024	100.00%
Unique queries	676,402	13.39%	722,689	19.42%
Repeat queries	4,373,813	86.56%	2,997,876	80.57%
Zero term queries	2,539	0.05%	459	0.01%
XML queries	39,151	0.77%	269,157	7.23%
English queries	4,983,919	98.64%	2,676,169	71.92%
Non-English queries	29,684	0.59%	775,698	20.85%

Table 2
Comparison of types of queries in Gnutella P2P network and Excite search engine

	Gnutella 2002 [3]	Gnutella 2003	Excite 1997 [17]	Excite 1999 [2]
Unique queries	13.39%	19.42%	57%	60.3%
Repeat queries	86.56%	80.57%	43%	39.7%
Zero term queries	0.05%	0.01%	N/A	N/A

in July 2002, increasing to 25.72 million in July 2003. The total number of Internet users in China also increased from 45.80 million to 68.00 million during this period. It should be noted that the number of broadband users has increased from 2.0 million to 9.8 million, an increase of 390%. The increase in the number of non-English queries is probably attributable to the drastic increase in broadband users in China, since the increase in bandwidth has provided an incentive for users to participate in file sharing, especially the sharing of large files such as movies.

In Table 3, it can be seen that there is a substantial increase in the number of terms per query between Gnutella 2002 and Gnutella 2003. The median number of terms per query has been increased from three in Gnutella 2002 to four in Gnutella 2003. The average number of terms per query has increased from 3.74 in Gnutella 2002 to 5.21 in Gnutella 2003. Unlike the Gnutella studies, the Excite studies did not show any significant changes in the number of terms per query. The median number of terms per query remains at two and the average number of terms per query has slightly increased, by 0.05, from 2.21 to 2.4 between Excite 1997 and Excite 1999. Figure 5 shows the distribution of the number of terms in Gnutella 2002, Gnutella 2003, Excite 1997 and Excite 1999.

In our previous study [3], we have shown that Gnutella queries are longer than Excite queries. The present study shows that the length of queries in Gnutella has further increased. The length of queries in Gnutella is now more than double that in Excite. Gnutella users are usually looking for a specific file. They also learn that the more specific the query is, the better the searching result will be. The Gnutella queries usually include the formats of the target files and the complete name of the media files, such as titles of songs, movies, and computer software. Names of movies can be as long as 10 terms or more. Besides, metadata, such as title, artist, album, and genre, can be included in the queries. For these reasons, the length of Gnutella queries is longer and longer. On the contrary, WWW users may have specific information to seek or have their information needs in mind but may not know exactly what the target documents will be. Therefore, more queries in the Web search engines are general at the beginning of a session and refined gradually. Besides, the file format is always in HTML format with other media files enclosed. As a result, the length of Excite queries remains at about two terms.

The content available in P2P networks usually has less variety in comparison with the content available on the Web [4]. P2P networks are usually more specific

Table 3
Comparison of number of terms per query in Gnutella P2P network and Excite search engine

	Gnutella 2002 [3]	Gnutella 2003	Excite 1997 [17]	Excite 1999 [2]
Median number of terms per query	3	4	2	2
Average number of terms per query	3.74	5.21	2.21	2.4

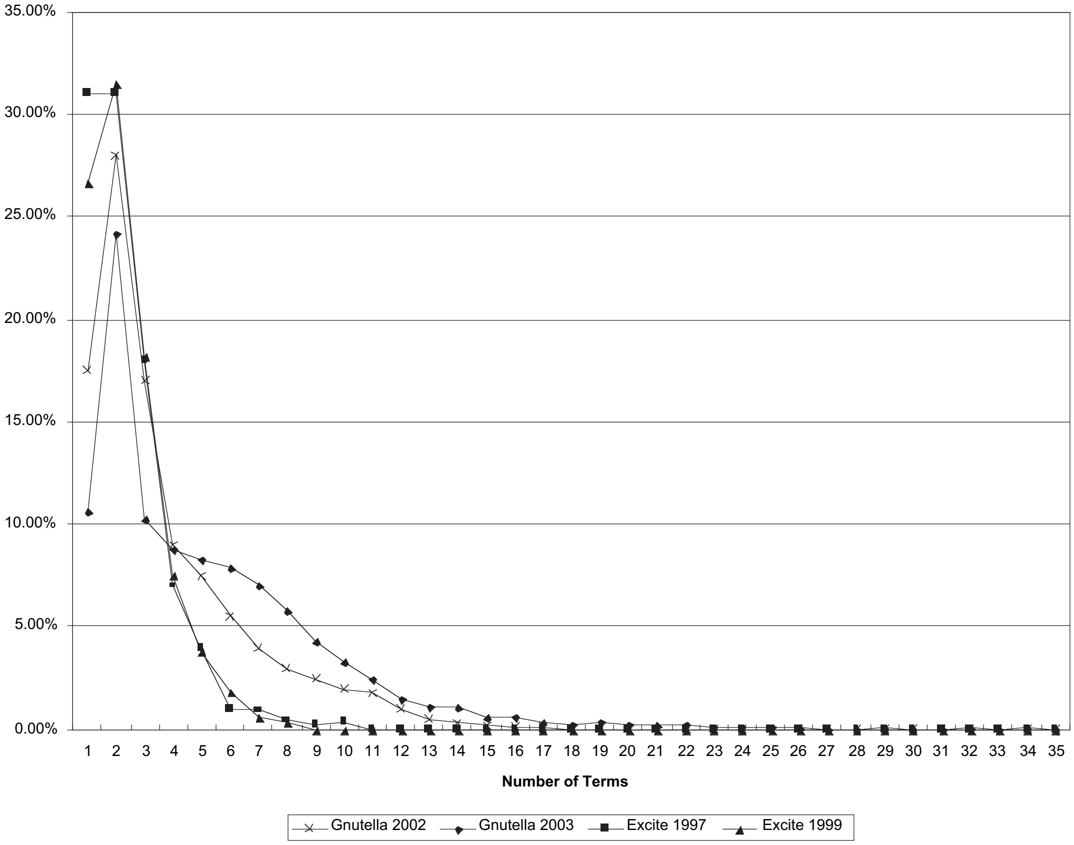


Fig. 5. Comparison of the distribution of the number of terms in Gnutella P2P network and Excite search engine.

to serve particular interests, such as entertainment and computer software. Other topics, such as news, medicine, business, and shopping, are not available in P2P networks but they are very common on the Web.

3.2. Changes in terms used in Gnutella queries

We have also investigated the terms being used in Gnutella queries and compared them with those in Excite queries. Terms being used in queries help us to understand the topics of interest to Gnutella users. Table 4 shows the top 15 queries in Gnutella 2002 and Gnutella 2003. It is interesting to note that queries associated with computer systems and related files that did not appear in Gnutella 2002 are now among the top 15 queries in Gnutella 2003. For example, 'wallpaper' is a file that is used as background for desktop or laptop computers. Portable document format (pdf) is a file format that preserves the original layout and content of documents that can be generated by several different computer softwares and is widely used on the Internet. 'Linux' is a popular operating

Table 4
Top 15 queries in Gnutella 2002 and Gnutella 2003

Gnutella 2002			Gnutella 2003		
Rank	Query	%	Rank	Query	%
1	divx	0.23	1	harry potter order phoenix	0.21
2	qwerty jpg	0.17	2	amber	0.20
3	porn	0.13	3	rm	0.17
4	eminem	0.13	4	WALLPAPER	0.15
5	techno mp3	0.12	5	rpg	0.15
6	divx avi	0.11	6	LINUX	0.13
7	porn mpg	0.10	7	SPEECH	0.13
8	spiderman	0.09	8	gangbang girls	0.13
9	chris isaak	0.09	9	MOVIE QUOTE	0.13
10	return to me	0.08	10	mo rmvb	0.13
11	joey gian	0.08	11	tom petty	0.09
12	Nelly	0.08	12	PDF	0.09
13	Sex	0.07	13	amber stories doc	0.08
14	aqua mp3	0.07	14	Robotech rpg adventurese rdf accelerated training program pdf	0.08
15	minority report	0.07	15	Two weeks notice avi	0.08

system used in PCs. In Gnutella 2002, query content is related to current entertainment and sexuality, such as recently released movies and songs, names of popular artists, and pornography. In Gnutella 2003, content related to current entertainment is still popular in the top 15 queries but content related to sexuality appears only once in the top 15 queries. This is an indication that the topics of interest have shifted from entertainment and sexuality to computer systems and entertainment. This pattern is similar to the Excite studies reported by Wolfram et al. [6] and Spink et al. [2] showing that topics of interest have shifted from entertainment and sexuality to e-commerce, people and computers.

The most frequent terms in all queries and in unique queries of Gnutella are also investigated. The results are presented in Table 5 and Table 6. Among the top 50 terms in all queries and in unique queries, file formats and terms related to sexuality and names of movies, songs and artists are the most popular. Most of the terms that appear in the top 50 terms in all queries of Gnutella 2003 but do not appear in the top 50 terms in all queries of Gnutella 2003 are those with current content, such as the names of movies or songs that are popular during the period of study, rather than changes in topics of interest.

4. Conclusion

There have been significantly fewer studies of information behavior on P2P networks in comparison with those on WWW search engines. However, there are differences between information behavior on P2P and on WWW search engines. In our previous work, we have reported that there are relatively more repeat queries in P2P queries and the length of queries in P2P is longer than those in WWW search engines. In this study, we find that there are more changes in the percentage of unique and repeat queries in Gnutella P2P networks than in Excite WWW search engines. The number of unique queries in P2P has increased and the number of repeat queries in P2P has decreased. However, the number of repeat queries in P2P is still greater than in WWW search engines. This is due to the random connectivity of peers in P2P. The length of queries in P2P has been increased while there is almost no change in the length of queries in WWW search engines. This is because the P2P users usually have a specific file in mind that is represented by the names of movies or songs while the WWW users may search for more general information that can be represented by keywords.

In our analysis of the terms being used in the queries, we find that the topics of interest among P2P users have shifted from entertainment and sexuality to computers

Table 5
Top 50 terms in all queries of Gnutella 2002 and Gnutella 2003 (after removing common terms without content)

	Gnutella 2002	Gnutella 2003		Gnutella 2002	Gnutella 2003
Rank	Term	Term	Rank	Term	Term
1	mp3	mp3	26	movie	assemblage*
2	urn:#	mpg	27	trek#	porn
3	avi	avi	28	episode#	pdf
4	mpg	rm*	29	big#	remix
5	zip	wmv*	30	men#	ram*
6	mpeg	mpeg	31	fuck#	live
7	jpg#	queen*	32	gay#	album*
8	you	symphony*	33	anal#	die*
9	xxx	you	34	young#	mix
10	sex	i*	35	john#	movie
11	porn	s*	36	remix	teen
12	star#	harry*	37	boys#	world*
13	divx#	potter*	38	mix	everything*
14	love	zip	39	your	civilization*
15	me	dat*	40	all#	life*
16	black#	asf	41	man#	rar*
17	my	xxx	42	new#	cd*
18	teen	me	43	german#	full
19	asf	girl	44	wars#	your
20	girl	Sex	45	dj#	but*
21	full	my	46	eminem#	walking*
22	red#	order*	47	dvd#	txt*
23	live	phoenix*	48	time#	wounded*
24	hot#	avi	49	pdf	lesbian*
25	girls#	love	50	bangbus#	xvid*

Notes:
# – terms that are in the top 50 terms in all queries of Gnutella 2002 but not in those of Gnutella 2003.
* – terms that are in the top 50 terms in all queries of Gnutella 2003 but not in those of Gnutella 2002.

and entertainment while the topics of interest among WWW users have shifted from entertainment and sexuality to e-commerce, people and computers.

It is obvious that the information behavior of P2P users has been changing as the P2P technology becomes more mature and Internet users become more familiar with it. It will be encouraging to see if P2P users begin to use this technology for information sharing in business applications in addition to leisure purposes for their everyday life information seeking as we can see on the WWW. The information behavior of P2P users may be extended to occupation/school information seeking.

As stated by Wolfram et al. [6], longitudinal studies of Web searching provide valuable insights into how public searching is evolving, changing, and moving in certain directions. These insights support Web design and public policy decisions. In this study, we have also conducted a longitudinal study of P2P searching and compared it with a longitudinal Web searching study. We can see similar changes and major differences between the two channels. Therefore, some of the Web design that used to accommodate the changes in Web searching can also be employed in the P2P environment. However, due to the difference between the technologies of P2P and Web, some changes are unique. For example, tools to support XML in P2P are becoming more popular. This is also reflected in the increase in XML queries in our study. Based on this observation, more advanced techniques in handling XML queries, documents, and metadata can be further explored and developed to support needs in the P2P environment. Besides, the increase in non-English queries implies that non-English information is becoming more

Table 6
Top 50 terms in all unique queries of Gnutella 2002 and Gnutella 2003 (after removing common terms without content)

	Gnutella 2002	Gnutella 2003		Gnutella 2002	Gnutella 2003
Rank	Term	Term	Rank	Term	Term
1	mp3	mp3	26	all	girls
2	urn:+	mpg	27	mix+	rmvb§
3	mpg	avi	28	young	big
4	avi	rm§	29	fuck	young
5	zip	wmv§	30	gay+	live
6	you	you	31	girls	drei§
7	jpg+	sex	32	full	full
8	mpeg	xxx	33	dj+	queen§
9	me	pdf	34	john+	frabezeichen§
10	love	porn	35	red+	black
11	my	s§	36	new+	your
12	sex	me	37	man+	dat§
13	porn	love	38	time	book§
14	asf	mpeg	39	pdf	lesbian§
15	girl	my	40	pussy	txt§
16	live	girl	41	soundtrack+	movie
17	teen	asf	42	boys+	hot
18	black	rpg§	43	get+	pussy
19	xxx	teen	44	anal	lolita§
20	divx+	die§	45	song+	it§
21	star+	zip	46	movie	time
22	your	anal	47	rock+	harry§
23	remix+	fuck	48	little+	all
24	big	robotech§	49	dance+	cum§
25	hot	one§	50	music+	civilization§

Notes:
+ – terms that are in the top 50 terms in all queries of Gnutella 2002 but not in those of Gnutella 2003.
§ – terms that are in the top 50 terms in all queries of Gnutella 2003 but not in those of Gnutella 2002.

popular in P2P as well as in WWW. Cross-lingual information retrieval and multilingual text processing techniques will be very helpful for the further development of P2P searching in the future.

References

[1] R. Schollmeier, A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In: *Proceedings of the First International Conference on Peer-to-Peer Computing* (IEEE Computer Society, Linkoping, 2001) 101–2.

[2] A. Spink, B.J. Jansen, D. Wolfram and T. Saracevic, From E-Sex to E-Commerce: Web search changes, *IEEE Computer* 35(3) (2002) 107–9.

[3] 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* 55(9) (2004) 783–93.

[4] K.Y. Chan and S.H. Kwok, *Information Seeking Behavior in Peer-to-Peer Networks: an Exploratory Study. Technical Report* (Department of Information and Systems Management, The Hong Kong University of Science and Technology, Hong Kong, 2003).

[5] 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 and Technology* 52(3) (2001) 226–34.

[6] D. Wolfram, A. Spink, B.J. Jansen, and T. Saracevic, Vox populi: the public searching of the Web, *Journal of the American Society for Information Science and Technology* 52(12) (2001) 1073–4.

[7] T.D. Wilson, Information behaviour: an interdisciplinary perspective, *Information Processing & Management* 33(4) (1997) 551–72.

[8] S.H. Kwok, File-sharing activities over BT networks:

pirated movies, *ACM Computers in Entertainment* 2(2) (2004) 11.

[9] R. Matei, A. Iamnitchi and P. Foster, Mapping the Gnutella network, *IEEE Internet Computing* 6(1) (2002) 50–7.

[10] C.C. Yang and A.Chung, A personal agent for Chinese Financial News on the Web, *Journal of the American Society for Information Science and Technology* 53(2) (2002) 186–96. [Special Issue on Web Research]

[11] C.C. Yang, J. Yen and H. Chen, Intelligent Internet searching agent based on hybrid simulated annealing, *Decision Support Systems* 28(3) (2000) 269–77. [Special Issue on Intelligent Agents and Digital Community]

[12] H. Chen, Y. Chung, M. Ramsey, and C.C. Yang, An intelligent personal spider (agent) for dynamic Internet/ Intranet searching, *Decision Support Systems* 23(1) (1998) 41–58.

[13] H. Chen, Y. Chung, M. Ramsey, and C.C. Yang, A smart itsy bitsy spider for the Web, *Journal of the American Society for Information Science* 49(7) (1998) 604–18. [Special Issue on Artificial Intelligence Techniques for Emerging Information Systems Applications]

[14] E.P. Markatos, Tracing a large-scale peer to peer system: an hour in the life of Gnutella. In: *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid* (CCGRID2002) (IEEE Computer Society, Berlin, 2004) 65–74.

[15] Gnutella, *The Gnutella Protocol Specification v0.4*. Available at: www9.limewire.com/developer/gnutella protocol_0.4.pdf (accessed 25 November, 2003).

[16] S.H. Kwok and K.Y. Chan, An enhanced Gnutella P2P protocol: a search perspective, *Journal of Interconnection Networks* 5(3) (2004) 267–78.

[17] B.J. Jansen, A. Spink and T. Saracevic, Real life, real users and real needs: a study and analysis of users queries on the Web, *Information Processing & Management* 36(2) (2000) 207–27.

[18] B.J. Jansen and U. Pooch, A review of Web searching studies and a framework for future research, *Journal of the American Society for Information Science and Technology* 52(3) (2001) 235–46.

[19] C. Yang and J. Luk, Automatic generation of English/ Chinese thesaurus based on a parallel corpus in law, *Journal of the American Society for Information Science and Technology* 54(7) (2003) 671–82. [Special Topic Issue on Web Retrieval & Mining: a Machine Learning Perspective]

[20] C. Yang and K.W. Li, Automatic construction of English/Chinese parallel corpora, *Journal of the American Society for Information Science and Technology* 54(8) (2003) 730–42.

[21] China Internet Network in Information Center (CNNIC), *Semiannual Survey Report on the Development of China's Internet*. Available at: www.cnnic.net.cn/ develst/repindex-e.shtml (accessed 29 November 2003).