2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.2.69

Cluster Analysis and Research of the Resource in KAD

Wang Chunhui[+], Chen Xingshu and Wu Qi

School of Computer Science, Sichuan University

Chengdu, China

Abstract- With the development of P2P file sharing in recent years, KAD network has been widely used. Although there are a large amount of resources in KAD, few of them can meet the users' demands. In order to find useful rescources in KAD, We need a systematic analysis of its characteristics. Firstly, describe the file score qualitatively, and cluster the resources according to the file size and the file score. Then find out the resources' features in the network. And finally use the analysis result to evaluate the search result of eMule client, make the conclusion that now search engine of eMule client cannot well satisfy the users' needs.

Keywords- K-mesns algorithm; P2P resources; KAD; cluster analysis

1. Introduction

With the development of P2P technologies, sorts of P2P softwares emerged. One of the features of P2P network is anonymity so that users can freely upload any resources. And currently most P2P resources will be automatically uploaded after being downloaded, which causes large amount of various resources distributed widely and spreading quickly over the P2P network. Therefore, users will inevitably find a lot of useless resources during their search. eMule is a P2P file-sharing software with numerous users. eMule v0.42b is formally embedded with the KAD which is a P2P network with neither server nor central node. While KAD network increases the download speed and provides the user with more download sources, it allows users to upload resources more freely. Since KAD network is lack of central node or server which can verify the uploaded resources, it contains a wide variety of illegal resources. Now multitudes of people have done research on KAD network. Jie Yu et.al[3] studied the ID repetition in KAD network and illustrated its influence to the network. Steiner et.al[2] systematically measured node distribution and its online time in KAD network, and finally they found the abnormality due to the modified KAD client joining to the network.

The paper applies the cluster analysis of data mining to analyzing the resources in KAD network. Now cluster analysis has been used in many industries and fields. In [11], the author uses it in traditional Chinese medical science to classify and diagnose disease. In [10], the cluster analysis plays a role in the analysis of Chinese mainland's investment environment. Besides, a P2P trust model based on cluster recommendation is proposed in [6]. This paper is to:

- Propose the standard of KAD network's resources meeting users' demands.
- Describe the resources' features in KAD network through cluster.
- Based to the analysis result, make a comprehensive evaluation to the search results of eMule client. The result shows that eMule client cannot meet users' demands well.

2. Background

[+] Corresponding author.
E-mail address: 61916613@qq.com

2.1. K-means Algorithm

The most common method of cluster analysis is K-means cluster[1] which is based on partition and proposed by MacQueen in 1967. K-means cluster is comparatively simple and has a relatively fast convergence rate. As a non-supervised learning algorithm, it is used to divide the given sample set into the cluster of specified number. Its calculation process is[7]:

- Determine the category number K.
- Choose k data points as the initial centroids(cluster centers).
- For each data point x in sample set, calculate the Euclid Distance from x to each centroid, and assign x to the closest centroid.
- Re-computer the centroid using the current cluster memberships.
- Repeat the two calculate processes mentioned above until it reaches the specified number of iteration or until it meets the requirement of iteration termination.

The prerequisite of iteration termination is usually criterion function convergence, and criterion function uses the sum of square error(SSE) whose definition is as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |$$

E represents the SSE of all objects in the data set; and p that is the point of the sample space means the given data object and m_i is the mean value of the cluster C_i . (Both p and m_i are multi-dimensional.)

2.2. KAD Network

KAD is a Kademlia-based [5][12] peer-to-peer DHT routing protocol implemented by several peer-to-peer applications such as Overnet [13], eMule [14], and aMule [15]. The two open–source projects eMule and aMule do have the largest number of simultaneously connected users since these clients connect to the eDonkey network, which is a very popular peer-to-peer system for file sharing. Recent versions of these clients implement the KAD protocol.

Similar to other DHTs like Chord [16],Can [17], or Pastry [4], each KAD node has a global identifier, referred to as KAD ID, which is 128 bit long and is randomly generated using a cryptographic hash function. We call a continuous subset of the total KAD ID space that contain all KAD peers whose KAD IDs agree in high order 8 bits a domain[5]. The index information of the files which contain the same keyword will be uploaded to the same domain, where peers's ID agrees at least in the first 8-bits with the keyword's. The file upload in KAD network contains such steps as:

- A file name will be segmented into several keywords according to the punctuation marks such as space.
- Calculate the HASH value of each keyword (key word ID) and the file (file ID).
- Find several nodes whose ID are similar to the keyword's. Release the index information <keyword ID, file ID > to these nodes.
- In the same way, find several nodes whose ID are similar to the file's, and release the indexing information <file ID, owner information> to these nodes.

One of the characteristics of KAD network is that the index information is stored on the nodes whose ID are similar to the index value. And the index information might only appear on the nodes that are in the same domain. Another feature is that if a file contains more keywords, its information will be uploaded to more domains in KAD network.

3. Resources Analysis

3.1. Resource Properties

For the files in KAD network, through distributed crawler we can find file name, file's HASH value, file size, file type, the number of file source (the number of file sharers), etc. The analysis shows that the basic feature of a file can be described by the length of file name, file size, file type and the number of file source.

Firstly, the name of a file reflects its subject. A longer name includes more keywords, which results in that the subjects are more limited. Secondly, file size and file type usually embody the file's usability. For example, a video file with the size of merely a few KB is often useless. Finally, the number of a file source is a symbol of its popularity and diffusance.

For the users, they hope they can find the file which, on the one hand, is shared by many people in the network, and on the other hand, has the same subject with the keywords they entered. In the search results, all file names must contain the keywords entered by users, so if a file's name includes more keywords, the subject of this file are more likely to deviate from what the users really want. Consequently, we can define the file that users want to find as one that has lots of sources and has a relatively short name. We give this definition a mathematical description as a score, and the higher score means that the users' demands are satisfied better. It is defined as follows:

$$Score_i = \frac{\max\{fnl_1, fnl_2, \ldots, fnl_n\} - fnl_i + 1}{\max\{sc_1, sc_2, \ldots, sc_n\} - sc_i + 1} + \frac{sc_i - \min\{sc_1, sc_2, \ldots, sc_n\} + 1}{fnl_i - \min\{fnl_1, fnl_2, \ldots, fnl_n\} + 1} \quad (1)$$

n is the file's number in search results; fnl_i is name length of the ith file in search results, and sc_i is the source number of the ith file in search results.

3.2. Resource Acquisition and Evaluation

This is a case study of the files with keywords "世界杯(world cup)" in KAD network. In order to find all these files, we crawl the domain where there is HASH value of keywords "世界杯(world cup)" by improved puppet client of eMule, send to each node in the domain a search request, and store the returned files into database after excluding repeating data. Eventually we get 706 files.

Using (1) to calculate the score of each file in the crawl result, we find that the files of high score are few and the scores are quite different. But the files of low score are just the opposite because:

- Crawl result shows that 85% of the files have only one source, which makes the low-score files have not very much difference.
- The score-calculating method, to some extent, highlights the excellent resource (those files with short name and lots of sources).

The analysis shows that a huge gap is existed between the file of highest score (the length of name is 7 and the source number is 86) and that of lowest score (the name length is 242 and the source number is 1). This demonstrates that the resources in KAD network cannot meet the users' demands very well. Fig. 1 displays the score of the files. On the abscissa is the files' score ranking in descending order whereas on the left vertical axis is the file score and on the right vertical axis is the number of the files under corresponding score.

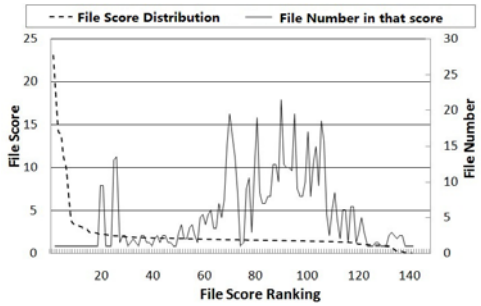

Fig.1. File score and the number of the files under corresponding score

3.3. The Results of Cluster and Analysis

In order to better find out characteristics of resources in KAD network, based on file score and file size as characteristic variable, we cluster the data by improved K-means algorithm[8][9]. In this way, the files are divided into three categories. TABLE I presents the features of these files.

The first category contains a total of 106 files among which 95% have obvious features: video files of about 200M, high score, and 97% of them are World Cup videos after analyzing their contents. Hence, we can conclude that these are files consistent with the subject "World Cup".

The second category has 511 files. These files are characterized by large fluctuations in size, and most of them have the score less than one point. Through content analysis, we find that most of them are illegal files or virus programs. These files are of different kinds and miscellaneous contents, which results in significant file size difference. And as the users want to upload their files to many domains of KAD network, they give the files with a relatively long name with many keywords. Besides, a part of eMule users download and share the files, so a few of the files in this category has some sources.

There are 89 files in the third category. They have few sources but higher score than that of the second category. The content analysis shows that these files are "World Cup" audio files, documents and pictures whose size is comparatively small. However, they get high score because their subjects are closely related to "World Cup". Thus we know that eMule users prefer to download and share video files of "World Cup".

Through the analysis above, we can described the files of the first category as the best resource that is relevant with the subject and satisfies the users' needs, the second category useless resource that is irrelevant with the subject and does not meet the users' demands, and the third category second-best resource that has relevance with the subject but is not popular like the first category.

TABLE I. FEATURE OF THREE CATEGORIES

Category	***Property***	***Maximum***	***Minimum***	***AVG***	***SD***
Best	File Size(M)	335	137	217	38
	File Score	23.1	2.1	3.1	3.3
Useless	File Size(M)	198	0.001	34	53
	File Score	3.2	0.02	0.8	0.7
Second-Best	File Size(M)	48	0.001	7	14
	File Score	14.5	1.2	1.7	1.2

1) Feature of File Size

File size, to a great extent, reflects the availability of the files which is related to a certain theme. The "World Cup" files contain various types, and among them tremendous size difference is existed. We can also see some characteristics about these files' size. According to the size, we rank these files with descending order. Fig. 2 compares the file size of the three categories. X-coordinate is the files' size ranking in their category, and Y-coordinate is file size.

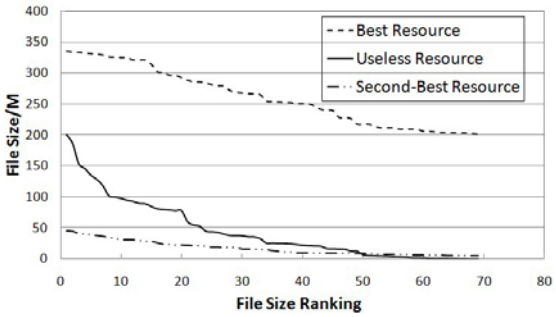

Fig.2. File size of three categories

From Fig. 2, we notice that the size of the useless resource varies quite widely, and it is lower than those of the second-best resource after the first 50 files. This is mainly because that the useless resource has different kinds of types, file size is highly correlated with file type, the similarity of file type is inversely proportional to the fluctuation of file size. That a number of very small files occur, to some extent, demonstrates that the useless resource has little usability.

2) Feature of File Score

The score can well illustrate the file name length and the number of file source. File score is directly proportional to its source number, and inversely proportional to name length. The files of higher score can better meet the users' demands. We alike rank the files with descending order based on their score. Fig. 3 is a comparison of the files of the three categories. X-coordinate is the files' score ranking in their category, and Y-coordinate is the file score.

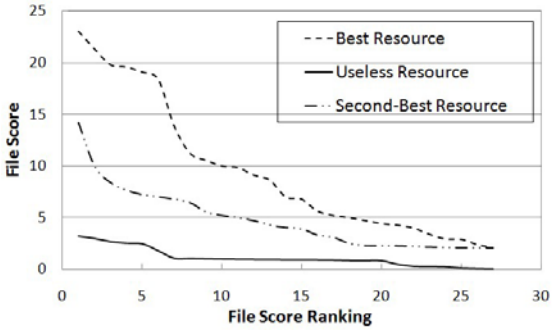

Fig.3. File score of three categories

Fig. 3 illustrates that the files of the best resource and the second-best resource have a great fluctuation in their score while the useless resource has less fluctuated and comparatively low score. The reason is that the best resource and the second-best resource contain some good files (name length is less than 15 byte and source number is more than 50) which have been clearly distinguished. As for those whose scores vary not much, they have very few sources and long name. Through analysis we find that there is a large amount of files that have only one source, which causes their low score and little fluctuation among the scores.

3) Feature of File Type

In TABLE II we can see the proportion of the files' type of these three categories. Most of the files in the best resource are video files of large size. However, the useless resource and the second-best resource contain different sorts of files, and the second-best resource apparently includes more picture, document and audio files than those in the useless resource. The useless resource is occupied by many files of pro type which are usually executable files, script files or some viruses. But pro type files are not expected to occur when we search keyword as "World Cup".

TABLE II. FILE TYPE PROPORTION OF THREE CATEGORIES

File Type	**Category**		
	Best	***Useless***	***Second-Best***
Unknown	8.6%	2.6%	8.4%
Video	82.8%	30.1%	12.8%
Pro	0	28.7%	0
Image	0	11.6%	22.3%
Doc	2.9%	8.7%	16.4%
Audio	0	8.9%	28.6%
Arc	5.7%	9.4%	12.5%

4) Summary

To conclude all the analyses mentioned above, we could describe the features of recourses in KAD network as follows:

- Lots of useless resources in the network. These resources, with various types, are mainly video, script and executable files, with their names containing dozens of keywords and their contents irrelevant with the keywords. Only few of such files have sharers.
- Files that match with users' requirements usually have numerous source and most of the file names have less than 10 keywords. File types and sizes are unified in spite of different subjects.
- Whatever the subject is, there exist abundant small-size files, most of which are of audio, picture and document types.

In order to manage the resources in P2P network, prevent the distribution of illegal resources, allow the users to find the resources they are needed in a fast and efficient way but maintain the privilege of anonymity, P2P softwares should enhance the capacity of filtering resources of different categories in terms of searching process and saving index information, which then both allows to shield illegal resources and stop their spread and allows users to seek out the resources they need. In the chapter to follow, the author of this paper is going to evaluate the searching results of eMule.

4. Evaluation of the Search Result

According to the cluster result of previous chapter, this chapter will evaluate the search result of eMule client. First of all, we classify the files that are searched by eMule client. TABLE III is these files' categories, from which we can see that the useless resource is at a high proportion. These files are mainly the videos and pictures of unhealthy contents as well as some very small files that are executable or of uncertain type.

TABLE III. THE CATEGORY PROPORTION OF EMULE SEARCH

	Category		
	Best	***Useless***	***Second-Best***
eMule Search	9.6%	84.5%	5.9%

Then we take a look at the file score of eMule search result. The highest score is 12.9, and the lowest 0.66. 88% of the files get a score lower than 4 among which 96% are the useless resource. Fig. 4 is CDF of the files' score. The analysis demonstrates that in eMule search result there are lots of useless resources, which cause that it is difficult for the users to find the exact resource they need. The reasons are that:

- Index information of many useless resources is stored in the nodes of network, these nodes may return many useless resources to searcher, and eMule stops searching when it receives 300 results, which decreases the probability of finding the best and the second-best resources.
- eMule client has no good mechanism to analyze the resources so that it shows the users all the resources which contain the keywords entered by them, despite that these resources' contents have no relation to the keywords.

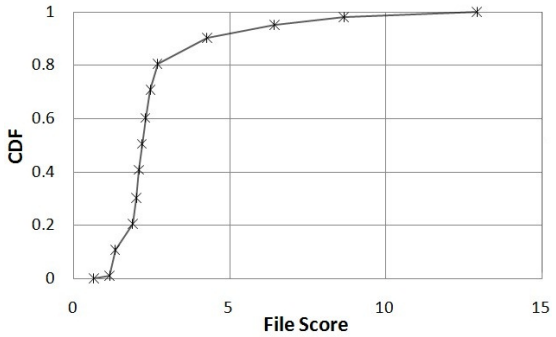

Fig.4. CDF of file score of eMule search

To sum up, the resources in eMule KAD network have various types and contents, but many of them include the keywords that are not perfectly relevant with the contents. Users who want to search the files they need must limit the file size and the source number. Although at current stage, eMule gives the users some reminders in the search interface, it still needs to make a great improvement for the users who are eager to find the resource they are interested in, because there are huge resources in KAD network.

5. Conclusions

This paper systematically and comprehensively analyzes the resources that contain certain information in a domain of KAD network, and summarize the characteristics of KAD resources, which will help the researchers to have a better understanding of KAD resources. Since eMule KAD search engine cannot well identify and censor the resources so that it presents the data without filtering. This inevitably leads users to get useless resources. Next the author will do some research on how to better provide the users with useful resources.

6. References

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkley Symposium Math. Stat. Prob, 1967, pp. 281 - 297.

[2] M. Steiner, E. W. Biersack and T. Ennajjary, "Actively monitoring peers in kad," In Proceedings of the 6th International Workshop on Peer-to-Peer Systems (IPTPS'07), 2007, pp. 1112-1117.

[3] Jie Yu, Chengfang Fang, Jia Xu, et al. "ID repetition in Kad," IEEE Ninth International Conference on Peer-to-Peer Computing, 2009, pp. 111 - 120.

[4] A. Rowstron, P. Druschel, "Pastry: Scalable, distributed object location and routing for large-scale Peer-to-peer systems," In Proceedings of Middleware Heidelberg Germany November. 2001.

[5] M. Steiner, T. En-Najjary and W. E. Biersack, "A global view of KAD," IMC'07: 2007 7th ACM SIGCOMM Internet Measurement Conference, 2007, pp. 117-122.

[6] Y. Lei, Q. Zhiguang and Z. Ting, "Peer-to-peer trust model based on cluster recommendation," Application Research Of Computers, vol. 27(04), 2010, pp. 1469-1472.

[7] H. Jiawei, M. Kamber, "Data Mining Concepts and Techniques," F. Ming, M. Xiaofeng, translate. Beijing: China Machine Press, 2007., in press.

[8] Z. Zhe, Z. Junxi, X. Huifeng, "Improved K-Means Clustering Algorithm," Congress on Image and Signal Processing (CISP '08), 2008, pp. 169 -172.

[9] L. Tian, W. Jianwen, "Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm," International Forum on Computer Science-Technology and Applications (IFCSTA '09), 2009, pp. 76-79.

[10] F. De-cheng, P. Xia, “Appraisal and Cluster Analysis on Regional Investment Climate of Mainland China,” 2006 International Conference on Management Science and Engineering (ICMSE '06), 2006, pp. 1112-1117.
[11] H. Qingyong, W. Jie, Yunling Zhang, et al, “Cluster Analysis on Symptoms and Signs of Traditional Chinese Medicine in 815 Patients with Unstable Angina,” Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '09), 2009, pp. 435-439.
[12] P. Maymounkov and D. Mazieres, “Kademlia: A Peer-to-peer informatiion system based on the XOR metric,” In Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS), 2002, pp. 53–65.
[13] Overnet [EB/OL]. http://www.overnet.org/.
[14] E-Mule[EB/OL]. http://www.emule-project.net/.
[15] A-Mule[EB/OL]. http://www.amule.org/.
[16] I. Stoica, R. Morris, D. Karger, et al, “Chord: A scalable Peer-to-peer lookup service for Internet applications,” In Proceedings of SIGCOMM, 2001, pp. 149–160.
[17] S. Ratnasamy, M. Handley, R. Karp, et al, “A scalable content-addressable network,” In Proc. ACM SIGCOMM, 2001.