A P2P Network Model Based on Hierarchical Interest Clustering Algorithm

Fanbin Meng [1], Lei Ding [1], Sheng Peng [1] and Guangxue Yue [2]
1. School of Information Science and Engineering, Jishou University, Jishou, China
2. School of mathematics and computer, Jiaxin University, Jiaxin, Zhejiang, China
Email: {mfb00368, dinglei_39}@yahoo.com.cn

***Abstract*—A hierarchical clustering P2P network model based on user interest is presented in this paper to improve search efficiency. The routers in the network are abstracted as the activated peer node, and the activated routing table is constructed to depict the character of the interest cluster. The P2P network is divided into different subnets according into their interest character, and the corresponding topological structure is build. The corresponding search strategy is proposed based on the above mention method. The simulation results show that compared with the traditional algorithm the interest cluster method proposed in this paper can form cluster more rapidly, and gain the appropriate resources faster.**

***Index Terms*—interest cluster, P2P, activated peer node, search strategy**

I. INTRODUCTION

The P2P network can be divided into structured and unstructured P2P system according to the P2P overlay network [1]. Compared with the structured P2P system, the unstructured P2P system has more loose resource organization and management, and has more random topological structure. Furthermore, the unstructured P2P system has a relative low searching efficiency compared with the structured P2P system. So the searching algorithm of unstructured P2P system has been the research hotspot in recent years.

At present unstructured search algorithm in general can be divided into two categories: blind search and heuristic search. Blind search algorithm spreads the query information through the network, and spread the query information to each node continually. Blind search algorithm uses the flooding mode or randomly selected node to retransmit the query search for desired resources throughout the network. For example, the directed BFS method presented in [2] makes each node maintain and count the data，then the querying node will choose some neighbor node to relay the query message, which has the most returned results or Minimum delay. To decrease the network consumption, Lv Q et al. propose a query algorithm based on multiple random walks that resolves queries almost as quickly as Gnutella's flooding method while reducing the network traffic by two orders of magnitude in many cases [3]. Chawathe Y et al. propose several modifications to Gnutella's design that dynamically adapt the overlay topology and the search algorithms in order to accommodate the natural heterogeneity present in most peer-to-peer systems [4]. Chen ZG et al. designs a novel Super-node Overlay Based on Information Exchange called SOBIE. Differing from current structured and unstructured, or meshed and tree-like P2P overlay, the SOBIE is a whole new structure to improve the efficiency of searching in the P2P network [5]. Lo V et al. describe three generic supernode selection protocols we have developed for peer-to-peer environments: a label-based scheme for structured overlay networks, a distributed protocol for coordinate based overlay networks, and a negotiation protocol for unstructured overlays. The integrated approach to the supernode selection problem can benefit the peer-to-peer community through cross-fertilization of ideas and sharing of protocols [6].

The blind search algorithm has more network consumption than the heuristic search algorithm. Now the heuristic search algorithm has been the research hotspot in recent years.

Heuristic search algorithm makes use of the known information to assist in the search of resources. Being more intelligent relatively, heuristic search algorithm is more worthy to be paid attention to in the unstructured network search field.

There are some typical heuristic search algorithm, such as the Mobile agent method, the content caching method, the heuristic flooding search algorithm, and the interest-based query algorithm. For example, Zhu YW et al. present GES (Gnutella with Efficient Search) to improve search performance. The key idea is that GES uses a distributed topology adaptation algorithm to organize semantically relevant nodes into same semantic groups by using the notion of node vector. Given a query, GES employs an efficient search protocol to direct the query to the most relevant semantic groups for answers, thereby achieving high recall with probing only a small fraction of nodes [7]. Tsoumakos D et al. present an overview of several search methods for unstructured peer-to-peer networks. Popular file-sharing applications, through which enormous amounts of data are daily exchanged, operate on such networks [8]. FENG GF et al. present an adaptive, bandwidth-efficient and easily maintained search algorithm for unstructured P2P systems, namely PeerRank. The scheme utilizes the feedback from

doi:10.4304/jsw.8.5.1262-1267

previous searches to probabilistically guide future ones. In addition, an effective caching and indexing mechanism is introduced, which remarkably enforces the search performance [9].

Among all kinds of heuristic search algorithm, the interest-based query algorithm derived from the idea, which the users have the same interest in the past may continue to have the same interest in the future [10], has become a research hot spot because the ideal accord with the actual situation in realistic life.

Nowadays interest-based query algorithm has achieved large success, for example, Sripanidkulchai et al. proposed a shortcut method based on interest [11], which added some special connection on the Gnutella overlay network and connected the nodes with the same interest together. The experimental results show that the algorithm can avoid a lot of flooding, and the scalability of the system has been greatly improved. The CAC-SPIRP method presented in [12] makes the peers with abundant content self-organize into a cluster where a query will be first routed to, and a response which satisfies the condition will may be most likely found in the cluster. So the CAC-SPIRP method significantly reduces the network bandwidth usage and the searching range. Furthermore, the method allows a client prefetch the indices of a few of entire files from a small group of peers with same interests, thus to minimize unnecessary queries and reduce query response time. The method presented in [13] raises an efficient search mechanism based on interest-group, in which all of the system resources are firstly described using resource description framework, and then the peers in the description system are self-organized into different interest-groups. Some rules are firstly set up, and resources are divided into different groups according to these rules in [14]. A semantic overlay network generation algorithm is proposed in [15], which gradually forms distributed hierarchical P2P network architecture based on the node's stepwise clustering. The entire clustering process can be roughly divided into two stages, cluster in the local and fusion in the global. A distributed hierarchical the P2P structure (HP2PC) is proposed in [16]. The structure is built on a multi-layer overlay network made up of the neighbor nodes. The super node as its neighbor representative is recursively elected to participate in the construction of the higher level neighborhood. The nodes and their neighbor nodes work together to complete the clustering process of P2P within a certain level layer. The clustering process is divided into several different parts, and each part of it independently makes use of distributed K-means method to complete the clustering process. A query routing method is proposed in [17], which firstly discovers the response nodes for a query and remember them, then the information is used to search the nodes who might answer the query. A P2P search mechanism based on network topology and node interest preference is put forward in [18]. He S.S. et al. build a overlay network through computing the similarity of the nodes, that is to say, the search strategy is determined by information of query and the degree of connectivity between nodes [19]. An integrated framework using cluster-based hybrid network architecture is proposed to support collaborative virtual surgery. Multicast transmission is employed to transmit updated information among participants in order to reduce network latencies, while system consistency is maintained by an administrative server. Reliable multicast is implemented using distributed message acknowledgment based on cluster cooperation and sliding window technique. The robustness of the framework is guaranteed by the failure detection chain which enables smooth transition when participants join and leave the collaboration, including normal and involuntary leaving. Communication overhead is further reduced by implementing a number of management approaches such as computational policies and collaborative mechanisms [20]. Bai X et al. propose an interest-based clustering peer-to-peer Network (ICN) architecture. ICN uses a lot of Freenet mechanisms and is based on cache management. ICN is self-organizing, fully distributed, scalable, and logically hierarchical. In ICN, the upper level is bound by de Bruijn graph. Nodes in the lower level self-cluster based on interest [21].

Those algorithms mentioned above use the user's interest to build overlay logical structure and improve the efficiency of resource search. However, little attention has been focused on the role played by the router in the search process. Therefore, the focus of this paper is how to enhance the effect of the router to improve search efficiency.

The rest of this paper is organized as follows: in section 2, an interest-based clustering P2P overlay network structure is established, and a maintenance strategy is proposed. In section 3 we propose a corresponding search strategy. Then in section 4, we design a simulation experiment to demonstrate the efficiency of the P2P overlay network structure and the search strategy. Finally, we summarize our work in section 5.

II. Multilayer Overlay Structure based on Interest Clustering of the Active Node

The basic idea of this paper is to cluster the nodes according to their interest characteristics, and treat the router as an active node to form a P2P overlay network of multilayer interest-domain.

A. Metric of Node Similarity

Let S be the document set extracted from all of the shared documents in the whole peer-to-peer network, and F be the feature set constituted by all the feature of these documents in S, which contains f different features totally.

For the node n_i, the feature vector is $\boldsymbol{w}_i(w_{i,1}, w_{i,2}, \ldots, w_{i,f})$, where $w_{i,j}$ is the j-th dimensional feature value of the node n_i, and there is

$$w_{i,j} = \frac{D_{i,j}}{D} \log \frac{d}{1+d_j}, \tag{1}$$

where $D_{i,j}$ is the number of keywords only in the node n_i which contains the j-th feature, D is the number of

keywords of all the nodes, d is the total number of nodes, d_j is the number of the nodes contains the j-th feature. Then the similarity between the node n_i and the node n_j is

$$\cos(n_i, n_j) = \frac{\sum_f (w_{i,f} w_{j,f})}{\sqrt{\sum_f w_{i,f}^2} \sqrt{\sum_f w_{j,f}^2}}. \quad (2)$$

The above equation shows that the similarity is 1 if the feature vectors of two nodes are completely similar, and the value is 0 if the feature vectors of two nodes are completely different.

B. Establishment of the Interest Domain

The process of establishment of interest domain is realized as follows:

(1) A node joins an interest cluster through a detection process, and the node will make itself become a super node and create a new domain of interest if it is so-called the first node of the cluster, which means that cluster is not exists. Super node firstly select its nearest router in the backbone network as its active node, and then store the interest cluster's feature vector in that active node. Cluster feature vector $\boldsymbol{c}_i(c_{i,1}, c_{i,2}, \ldots, c_{i,f})$ is the average value of the feature vector of all the nodes in the cluster, such as

$$\boldsymbol{c}_i = \frac{1}{k} \sum_k \boldsymbol{w}_j, \quad (3)$$

where k means the number of the total nodes in the cluster, and $\boldsymbol{w}_j$ is the feature vector of the node j.

(2) If the new join node is not the first node in the cluster, which means it can find its active node in the backbone network, the node will sends a join application to the active node.

(3) The active node calculates the similarity between the cluster's feature vector and the new adding node's. The node will be added to the cluster if the similarity greater than the given threshold. It's worth noting that there may be several clusters which are similar to the new adding node according to the judgment standard, the new adding node only select the largest similarity cluster to join. The new adding node will contact with the super node of the cluster to complete its registration and calculate the similarity with the other nodes in the cluster when it is accepted by an active node. Finally, the new adding node will select the first N nodes as its neighbors according to the similarity between the new adding node and the nodes in the same cluster provided by the super node.

(4) If the number of clusters in the charge of the active node exceeds the upper limit, the new adding node will select the nearest node from the current active node in the backbone network to join. If there is only a active node in the P2P overlay network and the suitable cluster feature vector can't be found in the active node, the new adding node will be treat as a super node to create a new interest domain.

(5) If the new adding node can not find a suitable cluster in the current active node which it wants to add in, then it try another active node according to the direction of the neighbor table in the current active node. If the new adding node can't find one suitable cluster to join from all of the active nodes in the network, a new cluster will be created and the new adding node becomes the super node. The network topology structure is depicted as figure.1.

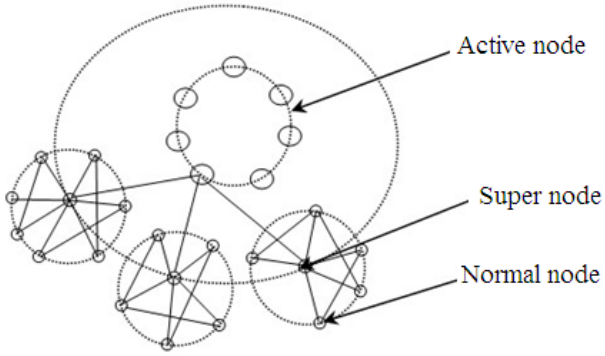

Figure 1. The Network Topology Structure

C. Information Provided by the Nodes

There are three types of nodes in the system, namely the active nodes, the super nodes and the normal nodes.

Active node is a router who plays a special role in the backbone network, it is also responsible for maintaining the key management information of some interest cluster, such as the super nodes, cluster feature vector etc besides the conventional routing functions. Furthermore, the active node also provides a neighbor active node table to facilitate the query of the cluster information, and the forwarding information table. The information provided by the active node is shown in table 1.

TABLE I.
The INFORMATION PROVIDED BY THE ACTIVE NODE

local resource table	The table of the super nodes belong to the active node	neighbor active node table	the forwarding information table

The super node is a representative node of a cluster, mainly maintains the related information of each member node in the cluster, such as the interest feature vector of the member node. In addition to this, the super node also maintains the identifier and address of other super nodes belong to the same active node. The information provided by the active node is shown in table 2.

TABLE II.
The INFORMATION PROVIDED BY THE SUPER NODE

backup super node table	local resource table	Neighbor super node table belong to the same active node	The table of the normal nodes belong to the super node	the forwarding information table

Normal node is mainly responsible for providing various types of shared resources, as well as the contents

of a neighbor table which are assigned by the super node. Furthermore, the normal node also provides the super node table and the forwarding information table. The information provided by the active node is shown in table 3.

TABLE III.
The INFORMATION PROVIDED BY THE NORMAL NODE

local resource table	The super node table which takes charge of the normal node	neighbor normal node table	the forwarding information table

D. Maintenance of the Interest Domain

The maintenance strategy of the nodes is given as follows:

(1) The election for the super node within a cluster will be held periodically. If a new node is elected as the super node, the older one will be degraded to a backup super node or normal node and its work will be taken over by the new super node. In addition to this, the backup node will stand up and hold election to produce a new super-node when the older super one exits.

(2) When a normal node exits, super node will broadcast the message to all nodes to make their neighbor table modified timely. The exiting node will inform the super node in advance under normal conditions, otherwise the super node will determine whether the node is normal or not by virtue of the timeout of periodic probe message response when the node exits abnormally.

(3) The super node will select a new node to be the backup node when it finds the backup node abnormal. The normal nodes will notify each other and re-elect a new super one when the super node and its backup node are abnormal simultaneously.

(4) If a new super node comes into being, it will contact with the active node associated with the cluster. In addition to this, the new super node will also notify its neighbor super nodes.

III. SEARCH STRATEGY

The search algorithm is given as follows:

(1) A query message of a normal node in a cluster will be sent to its neighbor nodes belong to the same cluster, and be forwarded to the super node if no successful reply returns.

(2) When a super node received a query message from the same cluster, it firstly searches the resources in the local storage, and will relay the query message to its neighbor super nodes if the related resources are not found. Finally, if the related resources are not found in all the super nodes and the corresponding normal nodes, the super node will hand in the query to the active node.

(3) The active node will firstly search the resources in the local storage if the query comes from its super nodes, and relay the query to its neighbor active nodes if no successful reply returns.

IV. ALGORITHM EXAMPLES AND ANALYSIS

The experiment document material is retrieved from homepages in www.Sohu.Com. The homepages in www.Sohu.Com are classified into about 5,000 documents involving 15 interests after pretreatment. The Peersim simulation platform is employed to realize the algorithm presented in this paper. There are 2000 nodes, and the out-degree of each node is set to be 4. Each node is randomly assigned about 1-4 interests, and a partial or the whole interest is stored in the node.

In this section, we in detail state the experimental result of the topological algorithm and searching algorithm. In order to more easy to analysis, the experimental result is made line charts.

A. Experement Results of the Topological Algorithm

The construction time of the interest cluster is employed to evaluate the performance of the topological algorithm presented in this paper. We repeated this experiment 10 times, and get the average value.

The comparison between the topological algorithm presented in this paper with the *k*-means algorithm in [13] is shown in Figure.2, which shows the varying time with the number of nodes joining the interest cluster. The abscissa means the signaling cycle of the node in Figure.2.

The conclusion can be derived from Figure.2 that compared with k-means algorithm the interest cluster method proposed in this paper can form cluster more rapidly.

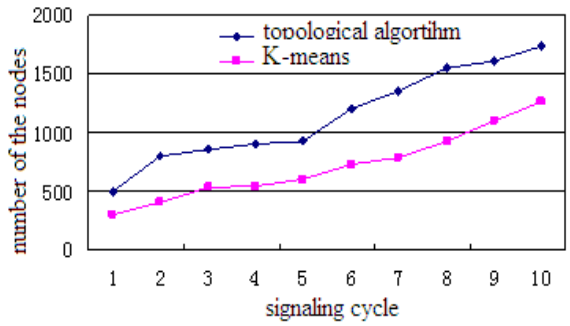

Figure2. Comparison between the topological algorihm with the *k*-means algorithm

B. Experement Results of the Searching Algorithm

In this experiment, the average query hops is employed to estimate the search performance. Here the average query hops is equal to the average passed hops before successful search.

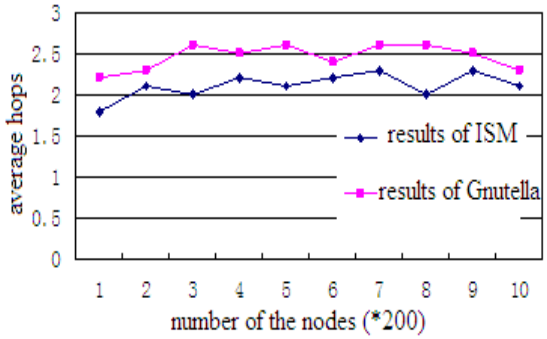

Figure 3. Comparison between ISM and Gnutella

Similarly, Figure 3. shows that the searching method (ISM) proposed in this paper can gain the appropriate resources faster than the Gnutella flooding searching algorithm.

C. Analysis

The topological algorithm presented in this paper takes the router as a active node, and constructs the active router table according to the characteristics of the interest clusters. The total P2P system is divided into different interest subdomains according to the active nodes which take charge of some interest clusters. The new adding node first searches the suitable active node to join, and participates the corresponding cluster taken charge of by the active node. In real life, people have the same interest in the past may continue to have the same interest in the future. So, the topological algorithm presented in this paper accords with the actual situation, and can form cluster more rapidly.

A query message of a normal node in a cluster will be sent to its neighbor nodes of the same cluster, and be forwarded to the super node if no successful reply returns. In reality, a node which returns a satisfied reply for a query will return a satisfied reply for next query, which accords with the idea which people have the same interest in the past may continue to have the same interest in the future. So, the searching algorithm will decrease abundant random searching process, and gains the appropriate resources faster than the Gnutella flooding searching algorithm.

V. CONCLUSION

The topological algorithm has been proposed according to the idea which people have the same interest in the past may continue to have the same interest in the future. In this paper, the router in backbone network is abstracted as the active node to participate in the construction of the overlay network and the search process. The focus of this paper is how to enhance the effect of the router to improve search efficiency.

Furthermore, the searching algorithm derived from the idea which returns a satisfied reply for a query will return a satisfied reply for next query decreases abundant random searching process, and gains the appropriate resources faster than the traditional Gnutella flooding searching algorithm.

Finally, the results of experiments show that the proposed method can make the user interest cluster constructed more quickly and improve the speed of the search to a certain extent.

ACKNOWLEDGMENT

This work was supported by grants from the Natural Science Foundation of Hunan Province No.07JJ6109 and the Science and Technology Plan of Hunan Province No.2010 GK3018.

REFERENCES

[1] Wang H H, Zhu Y W, Hu Y M. To unify structured and unstructured P2P systems [C]. Proceedings of the 19th IEEE international parallel and distributed processing symposium (IPDPS), Denver Colorado, 2005: 1-10.

[2] Yang B, Garcia-Molina H. Efficient search in Peer-to-Peer networks [C]. Proceedings of the 22nd IEEE international conference on distributed computing systems (ICDCS), Vienna, Austria, 2002 : 1-25.

[3] Lv Q, Cao P, Cohen E, Li K, and Shenker S. Search and replication in unstructured P2P networks [C]. Proceedings of the 16th ACM International Conference on Supercomputing, 2004: 1-20.

[4] Chawathe Y, Ratnasamy S. Breslau L, et al. Making gnutella-like P2P systems scalable [C]. Proceedings of the ACM SIGCOMM. Karlsruhe, Germany. 2003.

[5] Chen ZG, Liu JQ, Li D, Liu H. SOBIE: a novel super-node P2P overlay based on information exchange [J]. Journal of computers, 2009, 4 (9): 853-861.

[6] Lo V, Zhou DY, Liu YH, Gauthier-Dickey C, Li J. Scalable supernode selection in Peer-to-Peer overlay networks [C]. Proceedings of the 2005 Second International Workshop on Hot Topics in Peer-to-Peer Systems. 2005: 18-25.

[7] Zhu YW, Hu YM. Enhancing search performance on Gnutella-like P2P systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2006, 17 (12): 1482-1495.

[8] Tsoumakos D, Roussopoulos N. A Comparison of Peer-to-Peer Search Methods [C]. International workshop on the web and databases (WebDB), San Diego, 2003.

[9] FENG GF, MAO YC, LU SL, CHEN DX.软件学报，PeerRank:A Strategy for Resource Discovery in Unstructured P2P Systems [J]. Journal of Software, 2006, 17(5): 1098-1106.

[10] Azar Y, Fiat A, KafIin A R, et al. Spectral analysis of data [C]. Proceedings of ACM Symposium on theory of Computing(STOC), Hersonissos, Crete, Greece, 2001.

[11] Sripanidkulchai K, Maggs B, Zhang H. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems [C]. Proceedings of the IEEE infocom, San Francisco, US, 2003.

[12] Guo L, Jiang S, Xiao L, et al . Exploiting content localities for efficient search in P2P system [C]. Proceedings of the 18th International Symposium on Distributed Computing(DISC) , Amsterdam, Netherlands, 2004.

[13] Yang J, Zhong Y, Zhang S. An efficient interest-group based search mechanism in unstructured peer-to-peer networks [C]. Proceedings of the International Conference on Computer Networks and Mobile computing (ICCNMC), Shanghai, china, 2003.

[14] Cohen E, Fiat A, Kaplan H. A case for associative peer-to-peer overlays [C]. Proceedings of the workshop on hot topics in networks, New Jersey, US, 2002.

[15] Guo K, Liu Z. A New Efficient Hierarchical Distributed P2P Clustering Algorithm [C]. Proceedings of the fifth international conference on fuzzy systems and knowledge discovery, 2008: 352-355.

[16] Hammouda KM, Kamel MS. Hierarchically distributed Peer-to-Peer document clustering and cluster summarization [J]. IEEE transactions on knowledge and data engineering, 2009, 21 (5): 681-698.

[17] Tempich C, Staab S, Wranik A. REMINDIN: Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors [C]. Proceedings of WWW'2004, 2004.

[18] Liang WF, HUANG JH. Searching mechanism based on network topological and peer interest in P2P system [J]. Computer Engineering and Design, 2008, 29(6): 1316-

1318.
[19] He SS, GU NJ, TIAN ZX, XIE J. Search Algorithm Based on User Interest-proximity Measurement [J]. Journal of Chinese Computer Systems, 2008, 29(11): 2027-2030.
[20] Qin J, Choi K S, Poon W S, Heng P A. A framework using cluster-based hybrid network architecture for collaborative virtual surgery [J]. computer methods and programs in biomedicine, 2009, 96(3): 6205 – 216.
[21] Bai X, Liu S, Zhang P, et al. ICN: interest-based clustering network [C]. Proceedings of the 4th international Conference on Peer-to-Peer Computing (P2P), Zurich, Switzerland, 2004.

Meng Fanbin was born in 1964. He received the bachelor's degree in 1984 from Jishou University. Now he is an associate professor in school of information science and technology of Jishou Univerity. His research interests include computer network, etc.

Ding Lei was born in 1972. He received the Ph.D. degree in 2012 from central south university. Now he is an associate professor in school of information science and technology of Jishou University. His research interests include computer network, artificial intelligence and industrial process control.

Peng Sheng was born in 1974. He received a master's degree in 2009 from central south university. Now he is a lecturer in school of information science and Technology of Jishou University. His research interests include computer networks and intelligent systems.

Yue Guangxue was born in 1963. He received the Ph.D. degree in 2011 from Hunan University. Now he is a professor in Jiaxin University. His research interests include computer network and network security.