Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 8 (21): 3895-3902, 2008
ISSN 1812-5654

Measuring the Performance of Peer-to-Peer Systems with Social Networks Characteristics

A. Modarresi, A. Mamat, H. Ibrahim and N. Mustapha
Faculty of Computer Science and Information Technology, University Putra Malaysia, Malaysia

Abstract: In social system, people with similar interests gather and create a community. This structure organizes people effectively and makes sharing information easier. In sociology the behavior of such structures has been investigated for a long time. Fortunately this structure can be extended to Peer-to-Peer (P2P) systems. This is due to the fact that peers in P2P systems usually have few interests like people in the real world and they try to find other peers with similar interests. On the other hand, the structure of the underlying models in P2P has a direct effect on different aspect of such systems. In this study the performance related parameters of a P2P system with social network characteristics are measured by simulation. The result shows that using similar structure as same as real world inside a community produces better performance. In addition, flooding technique in such systems creates higher traffic than random structured model; however a simple controlled flooding can provide a satisfaction result.

Key words: Distributed systems, overlay network, community, model

INTRODUCTION

In theoretical point of view, P2P systems create a graph in a way that each node will be a vertex and each neighborhood relation between two nodes will be an edge of this graph. When no criterion is considered for choosing a neighbor, this graph will be a random graph; however two important factors (Chen *et al*., 2005) change this characteristic in P2P: 1) principal of limited interest which declares that each peer interests in some few contents of other peers and 2) spatial locality law. Since each node usually represents one user in the system, a P2P will be a group of users with different interests who try to find similar users. Such structure creates a social network. Moreover it has been shown (Barabási and Rékta, 1999) that in the real social network the probability of occurring a node with higher degree is very low. In other words, the higher the degree the least likely it is to occur. This relation is defined by power law distribution, i.e., $p(d) = d^{-k}$ where k>0 is the parameter of distribution, for degree of network nodes. The network model which has been defined with characteristics in Barabási and Rékta (1999) has a short characteristic path length and a large clustering coefficient as well as a degree distribution that approaches a power law. Characteristic path length is a global property which measures the separation between two vertices; whereas clustering coefficient is a local property which measures the cliquishness of a typical neighborhood.

When social network concepts are applied in P2P systems, designers can catch more information about a group of people who are using the network and the result is providing better services for the group according to their interests and needs. Orkut, Myspace and Winodws Live Space are some samples which use social network concepts.

As an example, we envision the scenario of sharing knowledge among researchers. Since each researcher has a limited number of interests, he can communicate with other researchers who work in the same area of interests. Because of many limitations like distance and resources researchers usually work with their colleagues in the same institute or college. Sometimes these connections can be extended to other places in order to get more cooperation. This behavior defines a social network with some dense clusters where these clusters are connected by few connections like Fig. 1. If one researcher is represented by one node, a P2P system is created which obeys social network characteristics and defines a community with a common interest. A dense cluster is usually a good

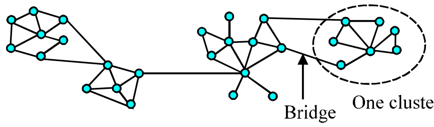

Fig. 1: Many similar dense clusters are connected by few connections and make a community

Corresponding Author: Amir Modarresi, Faculty of Computer Science and Information Technology, University Putra Malaysia, Jalan University, Serdang, 43400, Malaysia Tel: +60-17-2092620

source for common information in a P2P network. These clusters reduce unnecessary traffic in the network. The connections among clusters which are usually called shortcuts provide shorter path in order to find a proper source for rare information.

RELATED WORKS

Different structures and strategies have been introduced for P2P system for better performance and scalability.

Unstructured P2P systems like Gnutella usually create random graph. The most prominent searching technique in such systems is flooding. Although flooding is a simple and flexible technique, it suffers from lack of scalability. This technique produces huge network traffic and resource consumption especially in highly connected network. On the other hand a repeated query usually passes over nodes when loops exist in the structure of the network; therefore a bounding criterion like number of hops is used in order to stop searching. Such criteria limits search space in this method. Other techniques like modified BFS (Kalogeraki *et al*., 2002), iterative deepening (Yang and Garcia-Molina, 2002) and random walk (Lv *et al*., 2002) have better performance than flooding due to choose less neighbors based on some decision about neighbor selection.

Local routing indices have been introduced in order to select better neighbors based on contents of each neighbor (Crespo and Garcia-Molina, 2002). In such techniques the size of index tables is the main concern. Index tables grow by increasing of neighbors contents. On the other hand, if a node intends to decide about neighbors of a neighbor, the contents of those neighbors must also be indexed.

In super peer based systems like Napster, a special peer or group of peers, namely called super peers or super nodes, hold a global centralized index. All shared data by peers must be indexed in super peers. In contrast of fully centralized systems, after locating a particular peer, data transfer is done directly. Size of indices is still the main concern in these systems. Moreover leaving and joining peers need some modification in indices which is a costly task.

In structured systems the position of each node is tightly controlled. In such systems like CAN (Ratnasamy *et al*., 2001) and Chord (Stoica *et al*., 2001) if a piece of data exists in the system, its retrieval with some reasonable steps is guaranteed. It is commonly believed that administration cost is high in such systems, especially when the network changes dynamically.

Locality proximate clusters have been used to connect all peers with the same proximity in one cluster. Number of hop counts and time zone are some of criteria for detecting such proximity. Hu and Serviratne (2003) the general clusters have been introduced general clusters which supported unfixed number of clusters. Two kinds of links, local and global, connect each node to other nodes in their own cluster or nodes in other clusters. This clustering system doesn't concern about content of nodes; however, physical attributes are the main criteria for making clusters.

Crespo and Garcia-Molina (2005) created a Semantic Network Overly (SON) based on common characteristics in an unstructured model. Peers with the same contents are connected to each other and make a SON which is actually a semantic cluster. The whole system can be considered as sets of SONs with different interest. If a peer, for example, in SON S_1 searches contents unrelated to his group, finding proper peer is not always very efficient. If there is no connection between S_1 and the proper SON, flooding must be used.

Common interest is another criterion for making proper overlay. All peers with the same interest make a connection with each other, but locality of peers in one interest group has not been concerned In (Sripanidkulchai *et al*., 2003). All peers with the same interests are recognized after receiving many proper answers based on their interests (Chen *et al*., 2005). Such peers make shortcuts, a logical connection, to each other. After a while a group of peers with the same interests will be created and the richer peer in connection will be the leader of the group. Since this structure is based on unstructured system and receiving proper answer is in the range of the issued queries, we cannot expect that all peers with the same interests in the system are gathered in one group.

Shijie *et al*. (2006) have described community as the gregariousness in a P2P network. Each community is created by one or more peers that have several things in common. The main concern in this study was connectivity among peers in the communities. They have explained neither the criteria of creation nor size of each community. Khambatti *et al*. (2003) communities have been modeled like human communities and can be overlapped. For each peer three main groups of interest attributes have been considered, namely personal, claimed and private. Interests of each peer and communities in the system are defined as collections of those attribute values and peers whose attributes conform to a specific community will join it. Since 25 different attributes have been used in the model, finding a peer which has the same values for all of these attributes is not easy. That is why a peer may join in different communities with partial match in its attributes. Although the concept of the communities is as same as our work, in our model a shared ontology defines the whole environment and one community is part of the environment. There is also a bootstrapping node in each domain in order to preventing of node isolation. Present model also uses such nodes, but their main role is controlling sub communities. Haasea *et al.* (2004) used a shared ontology in unstructured P2P for peer clustering. Each peer advertises his expertise to all of his neighbors. Each neighbor can accept or reject this advertisement according to his own expertise. Expertise of each peer is identified by the contents of stored files. Since the ontology is used, a generic definition for the whole environment of the model is provided which is better than using some specific attributes.

Super peers have also been used for controlling peer clustering and storing global information about the system. Super peers are used in partially centralized model for indexing (Nejdl *et al.*, 2003). All peers who obey system-known specific rules can connect to a designated super peer. It creates a cluster that all peers have some common characteristics. Search in each cluster is done by flooding, but sending a query to just a group of peers will produce better performance. According to these rules, super peers who control common rules must create larger index; therefore they need more disk space and CPU power. Schlosser *et al.* (2002) instead of using rules, elements of ontology are used for indexing. In this structure each cluster is created based on indexed ontology which is similar to our method. All peers with the same attributes are indexed. Our model also uses super peers and elements of ontology for indexing, but instead of referring to each node in the cluster, super peers refer to the representative of that cluster which controls sub communities of a specific community. This will reduce the size of index to number of elements in ontology which is usually less than the number of peers in a large system and provide better scalability.

SUMMARY OF THE PROPOSED MODEL

The proposed model tries to implement social network concepts. In social network People usually make a social cluster based on their interests but in different size. Such clusters which are usually dense in connections are connected to each other by few paths. All of these clusters with similar characteristics create a community and these clusters make sub-communities. In each community: (1) each person must be reachable in reasonable steps and (2) each person must have some connections to others which are defined by clustering coefficient. With such characteristics some structures cannot show the behavior of social network due to the long average path among nodes like two dimensional lattice or lack of clustering like trees.

Providing a rigid structure increases administrative task burden; therefore it is tried to define the model as simple as possible that all nodes can contribute in it.

The model M has a set of peers P where: P = {p_1, p_2, .., p_n}. Each peer p_i can have d different direct neighbors. As a direct neighbor, p_k is one logical hop away from p_i which makes an overlay above physical network. Physical connection between p_i and p_k may not be a one hop connection.

A shared ontology O is used to define the environment of the system. Interests of peers are identified according to the ontology. O is stored in each peer in order to understand the structure of the environment. Based on ontology O many logical communities can be identified. Each community is populated by nodes with the same interests. Therefore all peers with the same interest can be identified by that community.

Contents of shared files in p_i identify the interest of p_i. The information about files is expressed by RDF statements comprises with shared ontology O. If p_i has different kinds of files which distinguish different interest, p_i can contribute in different community c_l, as a result, two communities can be connected to each other via p_i and these kinds of connections define shortcut among communities. If all communities are connected to each other all peers are reachable.

Inside each community, there are some peers who are rich in contents and connections as same as a knowledgeable person with good social relationship in social network. These peers are called hubs. Each hub defines a sub community inside the community.

Each community contains at least one member as a known member who is the representative of that community. This role is usually granted to the first peer who establishes a new community c_l and identified by r_l. We can consider a fellow f_l for representative r_l in community c_l, for reducing failure rate of the community when representative leaves the network. When the community is populated, r_l just refers to hubs inside the community. Since number of sub- communities inside each community is few, representatives do not need extra resources like CPU power or disk space to store or process this information. As the first known member of the community, representative can help other peers to settle in better place. Since in the real world, each community is a set of clusters or sub-communities and members of each cluster usually obey some kind of proximity, such a structure must be considered in the model. Good criteria to address the proximity can be number of hops or other metrics like IP address. While all peers in one community have similar interest, located peers with closer number of hops, it may provide closer distance among peers. Such configuration gives better response time for queries whose answers are in one community. In other word, locality of interest will be established in a better form in the community. This is done by introducing all hubs in the community to a peer who likes to join that community. The new peer can calculate his distance from each hub by sending a control message. The result will be a hub with a close distance as the first connection of the new peer. Peers according to their desire and/or capabilities can make more connections with other peers. This changes the structure of the model from tree-like structure to a graph which increases cluster coefficient of the system.

M also has a set of super peer SP where: SP = {sp_1, sp_2,..., sp_m} and m«n. Sp_i refers to the representatives of each community; therefore each community is identified in the system. Sp_i also stores the shared ontology of the system. This helps Sp_i to have a great view from all the system. As a bootstrap server, Sp_i can guide each new peer to a proper community just by knowing the interest of the peer. Since communities are mostly created based on the elements of the ontology and it is much less than number of peers in the system, the size of index in the Sp_i will be smaller than other super peers who work in semi structured model and need to index all peers or group of peers in the system. On the other hand, it provides the interconnectivity of whole system. Figure 2 shows an instantiate of the defined model based ACM ontology (ACM, 1998).

The complete definition of the model and its related algorithms has been introduced by Modarresi *et al.* (2008).

SIMULATION

A simulator is prepared to create a computer based community model to show the behavior of the system and in what extend they are close to a social network. An example as an instantiate of the model is explained. Based on this example proper dataset is provided.

A computer scientist regularly has to search publications or correct bibliographic meta data. A scenario which is explained here is community of researchers who share the bibliographic data via a peer-to-peer system. Such a scenario has also been expressed by Ahlborn *et al.* (2002) and Haase *et al.* (2004). The whole data environment can be defined by ACM ontology (ACM, 1998). Each community in the system is defined by an element of the ontology and represented by a representative node. Each community comprises of many sub communities or clusters which are gathered around a hub like Fig. 2.

Data set: A bibliographic file from DBLP server is used as a preliminary dataset (Ley, 1993). The contents of the file are categorized based on a simple syntactical method. If an ontology item from the shared ontology has been used in the title of a study in the file, it is assumed that the content of the file comprises the same interest with the ontology item. These pieces of data are presented in RDF statements which can be used by RDF query language. A sample of prepared data which is classified under information systems is like below:

```
<Publication rdf:about="dblp:persons/books/ph/Tomlin90">
 <title>Geographic Information Systems and Cartographic Modelling</title>
<acm:topicrdf:resource="http://daml.umbc.edu/ontologies/
classification#ACMTopic/Information_Systems"/>
 </Publication>
```

Title and topic of such a notation can be used in select and where-clause of a RDF-based query language like SPARQL (Prud, 2008).

Simulation setup: Thousand nodes are chosen for constructing the model. For each node a capacity for making connections with other peers based on power law distribution is considered. The first peer who joins the community is chosen as the representative of the community. Based on the definition of the model and number of sub-communities in each community, those

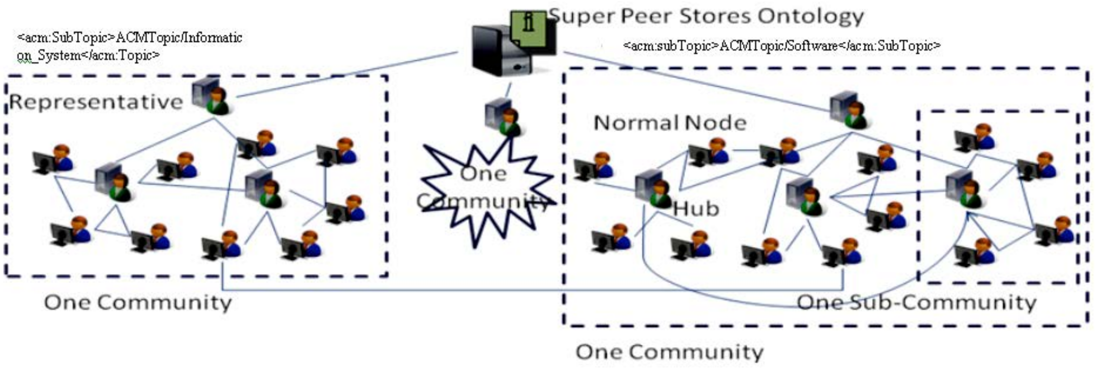
Super Peer Stores Ontology
<acm:SubTopic>ACMTopic/Information_System</acm:Topic>
<acm:subTopic>ACMTopic/Software</acm:SubTopic>
Representative
Normal Node
One Community
Hub
One Community
One Sub-Community
One Community

Fig. 2: An instantiate of the proposed model which uses ACM ontology

peers who are richer in connections are chosen as hubs. Hubs are normal peers with higher capacity for accepting connection. If all the connections of a particular hub have already been used another hub will be chosen by a new peer who wants to join to the community. Such a restriction in connection limitation has many reasons. First, it allows controlling the connection distribution in the system. Second, after all hubs are full, the new peer must connect to other normal peers. This mimics the behavior of joining a member to a community by another member. If the new peer has capacity more than one connection, other neighbors will be chosen randomly. First all the members inside the same sub-community are chosen because they may have shorter distance and then, if all peers cannot accept any more connections, the other peers from other sub-communities are chosen. These kinds of connections create potential bridges among sub-communities which make different sub-communities are connected without cooperation of the representative of a particular community. These kinds of connections increase the cluster coefficient of the model. Since the locality is important, such connections will be established when the target peers is rich in favor contents. Figure 3 shows frequency of nodes per connections created by simulator. In the figure the maximum number of connections that a peer can have is considered as 50.

In order to show the validity of the model with social network, we calculate cluster coefficient and path length for one community. If a query is not related to a community, the issuing peer may send the query directly to the super peer. The super peer will identify the representative of a proper community; therefore the whole path length of the model is gained by adding two extra hops to the calculated value. This is the maximum path length of the whole model by considering how many of queries can find proper answers outside their own community. Figure 4 shows the value of cluster coefficient when there is no hub (no specific structure inside the sommunity), 10, 50 and 100 hubs in the model with maximum connections of 10, 20 and 50 for peers conform to the power law distributions. Figure 5 shows the path length for the mentioned values.

By increasing number of hubs and connections, characteristic path length is decreased in Fig. 5. Both factors, number of hubs and connections, have a direct effect on this characteristic. In Fig. 4 when number of connections is high, peers have more capability for establishing connections. When one peer establishes connections with hubs and still has capability, it tries to make connections with other peers in the same community. The result is moving the model toward random network which explains less cluster coefficient when 50 connection is used.

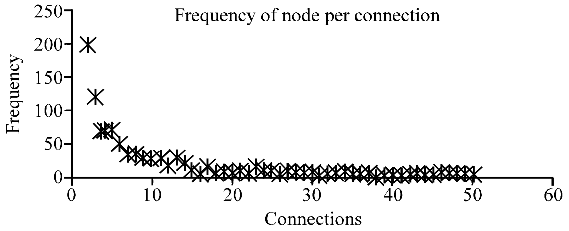

Fig. 3: Frequency of 1000 nodes based on power law distribution

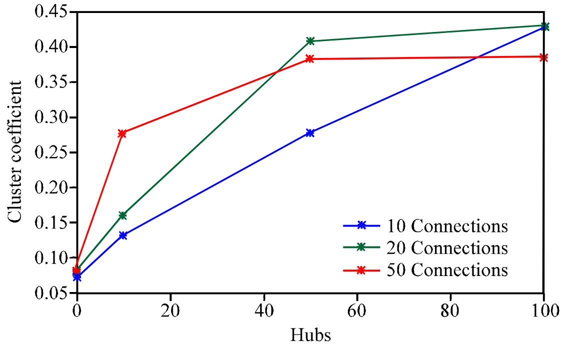

Fig. 4: The cluster coefficient of the model

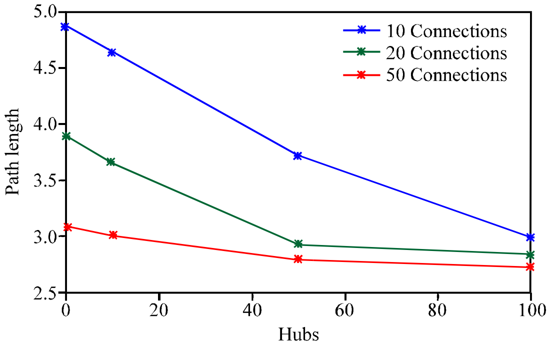

Fig. 5: The Path length of the model

A simulation is run for two different communities in the system. The first community shows information system in ACM ontology and the second one shows database management. For the first community 3970 items and for the second one 470 items exist in the bibliographic file. Each peer is loaded based on linear distribution with average 10 files. Seventy percent of files in each peer are related to the interest of each particular peer. Other 30% are chosen from two other interests, but the peer doesn't join to another related community; because the number of files is few. Eighty percent of all queries which are issued are related to the interest of peers; therefore the answers will be found in the same community. Other 20% of queries must be sent to their proper community (in this simulation to another community via the super peer). As

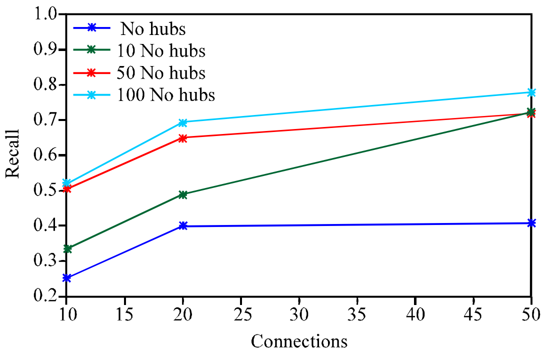

Fig. 6: Recall values per connections

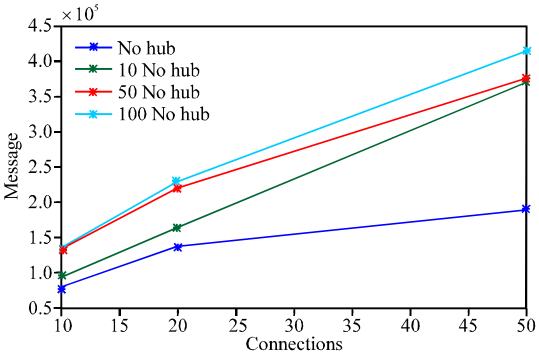

Fig. 8: Issued messages per connections

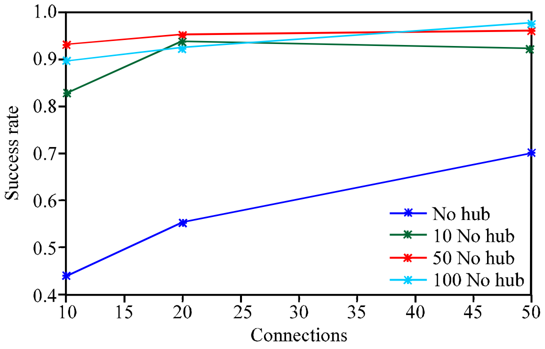

Fig. 7: Success rate per connections

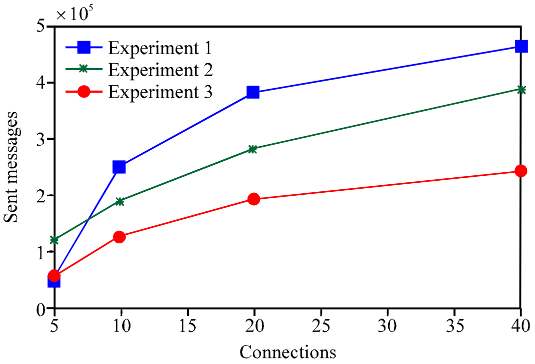

Fig. 9: Sent messages with different experiments

it was stated, 1000 peers are created in advance and are divided between two communities randomly. The routing algorithm is pure flooding with 5 hops.

Recall rate is defined as the number of retrieved related data per total related date. Figure 6 shows that by increasing number of connections, recall rate is increased. Since pure flooding is used, more connections mean more coverage during flooding. Success rate is also defined as the number of queries which receive at least one answer per total issued queries. Figure 7 shows that number of hubs and consequently cluster coefficient affects success rate. The best value is gained when all peers have a chance to establish a connection to hubs without unnecessary connections. Since hubs are good source of data and connections we could expect such results. Number of created messages is affected by number of connections which Fig. 8 shows.

It is observed that when no hub exists, number of created messages is fewer than other conditions. This is due to the fact that when a query is sent to a hub and that hub forwards the query to its neighbors, a lot of traffic is created. When there are many hubs, some of them may be connected to each other that make the situation worse than few numbers of hubs in a community. This problem can be solved by changing the routing algorithm slightly. Three other experiments are conducted as follow to show how a simple change in the flooding routing according to the structure of the model can reduce the traffic effectively. In experiment 3 each peer considers few simple criteria for choosing next peer.

Experiment 1: In this experiment a P2P random model is constructed and flooding algorithm is used for answering queries.

Experiment 2: Pure flooding is used in the proposed model with the mentioned conditions and 25 sub-communities in average. If a node poses a query which has other interest than its own, the node asks the super peer about proper community and its representative. Then the query is sent to that representative directly. In that designated community flooding will be done. Because just some part of network is searched, namely a community, it is expected that created messages are less than experiment 1.

Experiment 3: A control flooding algorithm is applied for answering queries in the proposed model like experiment

2. When a peer initiates a query, the query is sent to the representative of the community. If the query is not related to the peer's community and has different subject and interest, the initiating peer asks the super peer about proper community and its representative. Then, the query is sent to that representative directly. If a hub receives a query, it just sends the query to its neighbors excluding other hubs and the representative and if a normal node receives a query, it sends the query to its other normal neighbors.

The results of conducted experiments show that how a slightly changes in the routing algorithm can decrease created messages.

CONCLUSION

Using social network as an overlay for P2P systems can create better clustering and path length in comparison with random graph networks. Such overlays have a potential capability to answer queries. Measuring performance related parameters confirms this claim; however, when pure flooding is used in such systems number of created messages is increased. Considering some simple criteria based on network structure on flooding can produce a satisfaction result; therefore devising more efficient routing algorithms in a way that they can use the features of the overlay can be considered as future work of this job.

REFERENCES

ACM, 1998. The ACM computing classification system. New York, USA: Association for Computing Machinery. http://www.acm.org/class/1998.

Ahlborn, B., W. Nejdl and W. Siberski, 2002. OAI-P2P: A peer-to-peer network for open archives. Proceeding of the 31st International Conference on Parallel Processing Workshops, August 20-23, IEEE Computer Society, Vancouver, BC, Canada, pp: 462-468.

Barabási, A.L. and A. Réka, 1999. Emergence of scaling in random networks. J. Sci., 286: 509-512.

Chen, H., Z. Huang and Z. Gong, 2005. Efficient content locating in peer-to-peer systems. Proceeding of the 2005 IEEE International Conference on e-Business Eng. (ICEBE 2005), October 18-21, IEEE Computer Society, Beijing, China, pp: 253-256.

Crespo, A. and H. Garcia-Molina, 2002. Routing indices in peer to peer systems. Proceeding of the 22nd International Conference on Distributed Computer System (ICDCS'02), July 23-25, IEEE Computer Society, Vienna, Austria, pp: 23-32.

Crespo, A. and H. Garcia-Molina, 2005. Semantic overlay networks for P2P systems. Proceeding of the Agents and Peer-to-Peer Computing, 3rd International Workshop, July 19, Springer, New York, USA., pp: 1-13.

Haasea, P., B. Schnizlera, J. Broekstrab, M. Ehriga and F. van Harmelenb *et al*., 2004. Bibster-a semantics-based bibliographic peer-to-peer system. Web Semantics: Sci. Services Agents World Wide Web, 2: 99-103.

Hu, T.H. and A. Sereviratne, 2003. General clusters in peer-to-peer networks. Proceeding of the 11th IEEE International Conference on Network (ICON2003), September-1 October, IEEE Computer Society, Sydney, NSW, Australia, pp: 277-282.

Kalogeraki, V., D. Gunopulos and D. Zeinalipour-Yazti, 2002. A local search mechanism for peer-to-peer networks. Proceeding of the 2002 ACM International Conference on Information and Knowledge Management (CIKM), November 4-9, McLean, VA, USA., pp: 300-307.

Khambatti, M., R.K. Dong and P. Dasgupta, 2003. Structuring peer-to-peer networks using interest-based communities. Databases, Information System and Peer-to-Peer Computer 1st International Workshop, (DBISP2P), LNCS., 2944, September 7-8, Springer, Berlin Germany, pp: 48-63.

Ley, M., 1993. Digital bibliography and library project (DBLP). Germany Univ. Trier.

Lv, Q. P. Cao, E. Cohen, K. Li and S. Shenker, 2002. Search and replication in unstructured peer-to-peer networks. Proceeding of the 2002 International Conference on Supercomputing (ICS), June 22-26, ACM, New York, USA., pp: 84-95.

Modarresi, A., A. Mamat, H. Ibrahim and N. Mustapha, 2008. A community-based peer-to-peer model based on social networks. Int. J. Comput. Sci. Network Secur., 8: 272-277.

Nejdl, W., M. Wolpers, W. Siberski and C. Schmitz, 2003. Super peer-based routing and clustering strategies for RDF-based peer-to-peer networks. Proceeding of the 12th International World Wide Web Conference, May 20-24, ACM, Budapest, Hungary, pp: 536-543.

Prud, E., 2008. SPARQL Query Language for RDF. W3C. http://www.w3.org/TR/rdf-sparql-query.

Ratnasamy, S., P. Francis, M. Handley, R. Karp and S. Schenker, 2001. A scalable content-addressable network. Proceeding of the ACM SIGCOMM 2001 Conference on Application Technology, Architectures and Protocols for Computer Communication (SIGCOMM 2001), August 27-31, ACM, San Diego, CA, USA., pp: 161-172.

Schlosser, M., M. Sintek, S. Decker and W. Nejdl, 2002. HyperCuP-hypercubes, ontologies and efficient search on P2P networks. Proceeding of the Agents and Peer-to-Peer Computing 1st International Workshop, LNCS., 2530, July 15, Springer, Bologna, Italy, pp: 112-124.

Shijie, Z., Q. Zhiguang, Z. Xiaomei and Luo Xucheng, 2006. Interconnected peer-to-peer network: A community based scheme. Proceeding of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006), February 19-25, IEEE Computer Society, Guadeloupe, French Caribbean, pp: 108-108.

Sripanidkulchai, K., B.M. Maggs and H. Zhang, 2003. Efficient content location using interest-based locality in peer-to-peer systems. Proceeding of the IEEE INFOCOM 2003, The 22nd Annual Joint Conference IEEE Computer and Communications Society, 30 March-April 3, San Franciso, CA, USA., pp: 1-11.

Stoica, I., R.Morris, D. Karger, M.F. Kaashoek and H. Balakrishnan, 2001. Chord: A scalable peer-to-peer lookup service for internet applications. Proceeding of the ACM SIGCOMM 2001 Conference on Application, Technologies Architectures and Protocols for Computing Communications (SIGCOMM 2001), August 27-31, ACM, San Diego, CA, USA., pp: 149-160.

Yang, B. and H. Garcia-Molina, 2002. Improving search in peer-to-peer networks. Proceeding of the 22nd International Conference on Distributed Computer System, (ICDCS'02), July 2-5, IEEE Computer Society, Vienna, Austria, pp: 5-14.