An approach to minimize Topology Mismatch Problem in similarity -Aware Heterogeneous P2P Networks

Dr. B.Lalitha* C.Ravi Kishore Reddy**
Asst. Professor, CSE Department, JNTUACEA, India
**Asst. Professor, CSE Department, Tadipatri engineering college, India*
lalitha_balla@yahoo.co.in, ravikishore63@gmail.com

Abstract: The primary properties of the existing large scale peer-to-peer system include very high heterogeneity and dynamic nature of the participating peers in the network. Properties of peers such as peer session length, accessible bandwidth, storage space are highly varied with small set of peers controlling large part of the total system resources which results in low performance of the system. The nature of peers randomly joining and leaving the network causes the topology mismatch which drastically affects the network traffic. By taking along the advantages of similarity-aware overlays, the proposed system tries to reduce the effect of heterogeneous peers in unstructured peer-to-peer networks. The system is evaluated with proper simulations and the results show improvement in the total system capability.

Keywords- P2P, Unstructured, Overlays, Heterogeneity, Gnutella

I. INTRODUCTION

Peer-to-peer (P2P) is an approach to distributed data dissemination in which digital data is transferred between peer computers over the underlying Internet. P2P network forms distributed network architecture in which the participating peers share a part of their own computing resources which are essential for providing the Services and operations offered by the network and they are accessible by other peers. P2P networks can be broadly classified into Structured and Unstructured networks based on their structure. Structured networks such as Chord, Pastry are implemented using a Distributed Hash Table (DHT), which contain the details of the peers which are responsible for the data items in the network. These networks use identifier based searching and provide a guarantee of finding a data item. The drawback of structured networks is that it involves a large overhead in maintenance. In Unstructured networks the data is placed at random in the nodes and no node is responsible for any data item and searching is Object based. Due to the low overhead in maintenance the Unstructured P2P networks are widely used. An overlay network is a set of logical nodes and links that is formed above an already existing underlay network for providing or implementing a network service that is otherwise not available in the existing network. Overlay networks offer an alternative to modifying Internet protocols or routers, providing a quick and easy deployment path that lacks many of the technical and political hurdles of a router-level deployment. Overlays can take advantage of the large glut of processing, memory, and permanent storage available in commodity hardware to perform tasks that would ordinarily be well beyond the ability of a conventional router.

This paper mainly focuses on constructing an effective overlay topology to improve the search efficiency in a heterogeneous similarity aware overlay networks. The central focus will be on

1. Heterogeneous peers in the overlay network.
2. Topology mismatch problem

In P2P systems the characteristics of peers vary from one peer to other by a great extent. A small subset of peers controls most of the system resources while the rest have relatively a small amount of resources. The mismatch between the physical topologies and logical overlay forms the major factor that increases the overall response time for a search query. Mismatch problem also causes a large volume of unnecessary traffic in the P2P systems. Literature survey [1] shows that more than 75% of the P2P systems suffer from topology mismatch problem.

The remaining paper is structured as follows. Section

2 details the related work. In Section 3, we present the proposed system, and the method of topology construction and maintenance. In Section 4, we describe query searching algorithm. Section 5 presents the obtained simulation and numerical results. Finally, we conclude the paper with a summary in Section 6.

II. RELATED WORK

The major research focus in P2P for improving query search efficiency by designing good P2P overlay networks has provided some good overlay structures so as to speed up the searching process [2] [3] [4] [5] [6] [7] [8]. This paper mainly considers two structures which associate with the proposed system namely similarity-aware structure and location based structures.

The technical report [6] considers the similar contents among the peers for the formation of the similarity-aware overlay structure. Hai Jin et al [7] presented an overlay model that clusters peers which have similar content. In these overlay networks, similar peers are clustered together to form a SON (semantic overlay network). Queries are routed to the appropriate clusters which highly increases the chances of finding the related object and reducing the search load on nodes which has unrelated content.

Numerous solutions have been proposed to solve the topology mismatch between overlay topology and physical underlay network topology in P2P systems. LTM [3] built an efficient overlay in which nodes take the nearest nodes as their neighbors and removes the faraway nodes from their routing table according to the Round-Trip Time (RTT) information. SBO [5] organize the peers into red and white peers, where the white peers probe the distance with the other red peers and the red peers form the efficient overlay based on the probed distances. These findings tackle the topology mismatch problem, but considering physical locality is not enough. It is necessary to take the existing similarity of the common resources into account.

Yinglin Sun et al proposed a locality-aware group based semantic overlay [4] which incorporated the underlying locality into the semantic overlays. All the nodes took part in one or more semantic overlay(s). In each semantic overlay, nodes with close physical distance were organized into a group. There are some limitations for this approach. Each node needs to be categorized into several semantic classes firstly according to the categorical criterion. In generally, the categorical criterion is difficult to establishment. Furthermore, how to get the information of physical network distance does not mentioned in details.

The P2P systems may suffer from poor performance if they do not tackle the heterogeneity of peers and adapt the system according to the properties of the participating peers. The peers with lower processing capabilities or with lower network throughput are likely to affect the performance of the system. Even though heterogeneity is a challenge to the P2P system it can be seen as an opportunity that can be explored. As a result all the current P2P systems exploit this problem to their advantage. In majority of the P2P systems, peers are classified in two ways. The peers with high capability named *superpeers*, act as servers to the other lower capability peers. These *superpeers* form an autonomous topology within the system overlay and handle the important system functions. Normal peers maintain connections to the super-peers and act as clients to the *superpeers* [8]. KaZaA [9] uses super-peers (called super nodes) for client data indexing and helps in query search process.

Comparing with above algorithms, the proposed system considers the physical locality of the peers and the similarity of shared resources and at the same time taking advantage of the heterogeneous characters of nodes to form an overlay which provide better performance and efficiency in the object search process.

III. Proximity Based Similarity-Aware Unstructured P2P Networks

In this section the design of the proximity based similarity–aware overlay construction is presented. In short the design involves clustering of nodes into regions according to their physical proximity. Then the nodes in these regions are clustered into groups based on the similarity of the resources in the peers.

A. Overlay Design

Primarily the peers participating in the network are divided into two types namely *Super-peers* and *Normal-peers*. The Super-peers act as a centralized server to the Normal-peers by forwarding and replying the queries in behalf of the Normal-peers. A Normal-peer submits queries to Super- peers and also gets the results from them. The super peer construction implemented in this system exploits the heterogeneity of nodes by dispersing added responsibilities to higher-capacity Super-peer nodes. This categorization of nodes helps to implement load balancing as well as increase search effectiveness.

The topology is organized into various levels of peers as shown in Fig.1. The nodes in the each level are labeled as *Class1-SPs*, *class2-SPs* and *NPs*. The nodes which are in close proximity according to the network distance are clustered to form Regions, which are managed by *Class1-SPs*. The nodes in each region are clustered into groups based on the similarity of the resources in the peer. These regions are managed by *Class2-SPs*. The nodes which are controlling the regions i.e.*Class1-SPs* are connected to each other to form a pure P2P system. This overlay route helps to submit, forward and answering the queries.

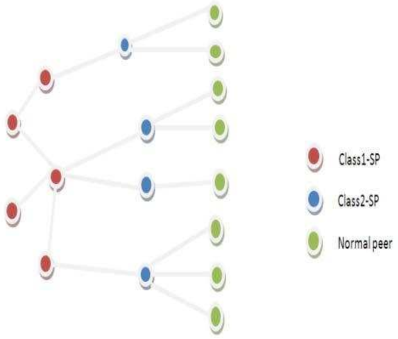


Figure1. Overlay Topology Structure

The topology construction in the system involves following phases: In the first phase node joining the network chooses its role as either a *Normal-peer (NP)* or a *Super-peer (SP)* based on the capabilities of the peer. Now a boot strapping node provides a list containing IPs of existing Class1- SPs. In the second phase the new comer resolves the closest region and applies to join it. In the third phase the Class1-SP will assign the new peer an appropriate group which has the most similarity of the shared resources with the new comer. In the fourth phase the new peer joins the group. The important decisions concerning the topology construction are as follows:

1) *Proximity Information Generation:*

To classify the nodes into clusters it is necessary to generate the proximity information. The network distance matrix can be efficiently represented by mapping its nodes to real geometric space. For simplicity the distance is represented by the *RTT (round trip time)* between two peers.

If *$\delta(p, pr)$* is the physical distance between node p and node pr, measured by the RTT values of two peers, which is define as follows:

$$\delta(p,\ pr)=\ \min\ \delta(p,\ pr)\ \forall\ p\ \varepsilon\ \{\text{Class1-SPs}\}$$

Peer p is assigned to region (pr) which has the minimum RTT.

2) *Super Peer Election*

Decision of which peers to act as super peers is resolved using the super peer election process. In this system a simple method involving that any peer can elect itself as the super peer if it has higher bandwidth, longer online times and superior processing power compared its counterparts.

3) *Similarity Function*

The nodes in the regions are clustered into groups according to the similarity of the shared resources. In this system *peerwords* are defined as the set of words which represent the character of shared resources in a peer. *Peerwordsimilarity (p, q)* is the degree of similarity between two peerwords of peer P and Q. The tuples are tokenized and the resulting vector representations are compared. The assignment of weights for the tokens in each tuple is crucial for effectiveness of similarity function.

The degree of similarity of two strings is denoted by similarity (a, b) which is defined as:

$$\text{Similarity }(a,\ b) = \textstyle\sum_{t\ \in\ a\ \cap b} W\ (a,\ t).\ W(s,\ t)$$

Where “W” is the weight function defined in terms of term frequency:

$$W\ (a,\ t) = \log\ (t.f_{a,\ t} + 1).\ \log\ (\ N\ /dft\ \ + 1)$$

here $t.f_{a,\ t}$ the frequency of t in a, N is the overall number of tuples and dft is number of tuples in which t appear

IV. Search Query Algorithm

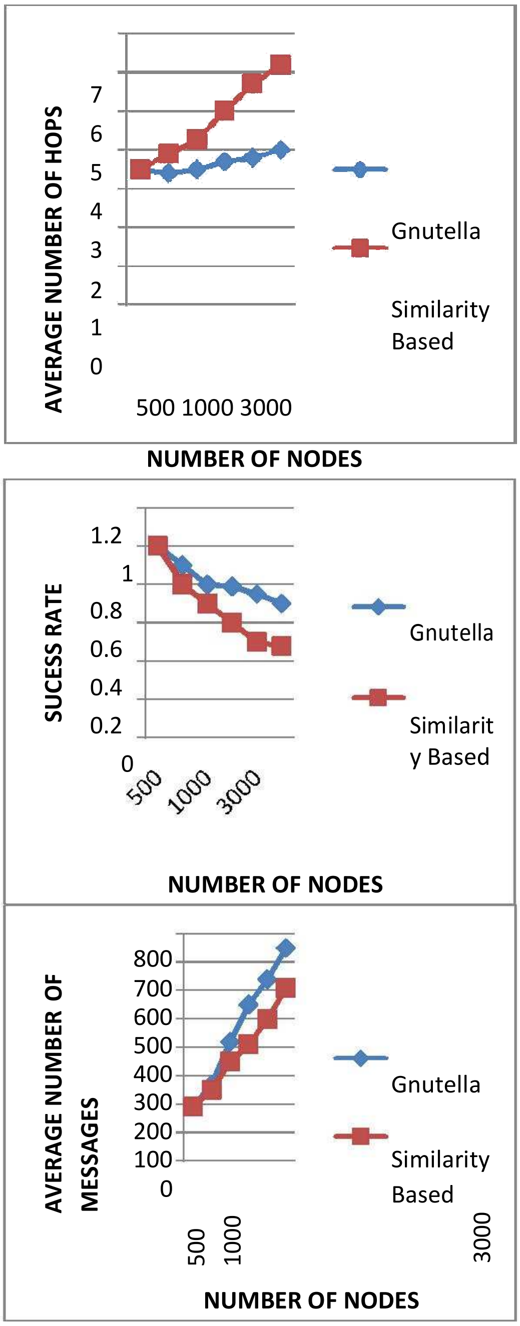
Nodes in the system maintains a routing table which contain the information of parent, parent backup, children and their *peerwords* along with other information. The query from a node is submitted to the superpeer it is connected. If the superpeer cannot answer the query, the query is forwarded to the upper layer superpeer. In the Class1-

SP layer the query is flooded as in a pure P2P system. When a Class1-SP receives a query message from another, the superpeer will determine which child to send based on its local information.

V. Simulation and Experimental Results

This section presents the system simulation setup along with the findings of the simulation. The system is implemented based on an open source P2P simulator. The system is simulated in two perspectives i.e. similarity clustering and proximity clustering. The findings are compared with the general Gnutella [10] environment. We evaluate the results based on the metrics of traffic cost, average number of hops and the successful queries.

The experimental setup involves about 500 to 5000 nodes. The degree of each superpeer to 7 and the number of Class1-SP is about 6 percent of the total size of the network. The simulation generates 500 ueries and each query is started at a random node.



Plot1. Similarity Gnutella Comparison

As depicted in Plot1, the success rate and the average number of hops in Gnutella is better compared to this system but the average number of messages transferred per query is quiet high in Gnutella. This shows the flooding mechanism in Gnutella is effective but produces too many redundant messages which waste the bandwidth and resources.

The second set simulation concentrates on the effect of proximity based clustering in the system. The simulation finds the physical path length of the

search by considering the average run time per query. The performance of the system is compared with the Gnutella network where the nodes will not consider the physical proximity of the other nodes. Table 1 displays the average searching process time and success rate in both the overlays.

The success rate of Gnutella is good but the average query processing time is the longest. This shows that the physical path length of search in Gnutella is too long. The average run time per query is greatly reduced in the system because it alleviates the mismatch problem and also improves the chances that a matching resource will be found quickly through similarity clustering.

Overlay	Average run time per query	Success rate
Similarity clustering	150.9	90%
Gnutella	750	98%

Table1: Comparison Gnutella vs. Similarity Overlay

VI. Conclusion

The proposed system is highly scalable, self organizing P2P network which solves the problem of heterogeneity and topology mismatch by clustering the nodes based on similarity of shared resources and proximity node clustering. The system reduces the redundant traffic generated unlike the other unstructured P2P systems which does not consider the node proximity information. Simulations and analysis verify that the model is realistic and effective.

VII. References

[1] Shen, G., Wang, Y., Xiong, Y., Zhao, B., Zhang, Z.: HPTP: Relieving the Tension between ISPs and P2P. In: IPTPS (2007)

[2] Wang Huijin, Lin Yongting, Cone: A Topology-Aware Structured P2P System with Proximity Neighbor Selection, International Conference on Future Generation Communication and Networking, 2007, pp.41-47

[3] Yunhao Liu, Li Xiao, Xiaomei Liu, Lionel M. Ni,Xiaodong Zhang, Location Awareness in Unstructured Peer-to-Peer Systems, IEEE Transactions on Parallel and Distributed Systems, Vol.16, No. 2, February 2005, pp. 163-174.

[4] Yinglin Sun, Liang Sun, Xiaohui Huang, Yu Lin, Resource Discovery in Locality-aware Group-based Semantic Overlay of Peer-to-Peer Networks, Proceedings of the 1st International Conference on Scalable Information Systems (INFOSCALE'06), May 2006.

[5] Yunhao Liu, Li Xiao, and Lionel M. Ni, Building a Scalable Bipartite P2P Overlay Network, IEEE Transactions on Parallel and Distributed Systems, Vol.18, No. 9, September 2007, pp. 1296-1306.

[6] Hung-Chang Hsiao, Hong-Wei Su. On Optimizing Overlay Topologies for Search in Unstructured Peer-To-Peer Networks. IEEE transactions on parallel and distributed systems, Vol. 23, No.5 , May 2012.

[7] Hai Jin, Xiaomin Ning, Hanhua Chen, Zuoning Yin, Efficient query routing for information retrieval in semantic overlays, Applied Computing 2006. 21st Annual ACM Symposium on Applied Computing, 2006, pp. 1669-1673.

[8] B. Yang and H. Garcia-Molina. Designing a super-peer network. *19th International Conference on Data Engineering*, pages 49-60, Bangalore, India, March 2003. IEEE Computer Society.

[9] J. Liang, R. Kumar, and K. Ross. The KaZaA overlay: A measurement study. *Computer Networks*, 50:842-858, April 2006.

[10] Gnutella http://rfc-gnutella.sourceforge.net/,2011