Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

# A PROXIMITY-AWARE INTEREST-CLUSTERED P2P FILE SHARING SYSTEM

**Dr.S. DHANALAKSHMI[1], R. ANUPRIYA[2]**
[1]*Prof & Head,* [2]*Research Scholar*
*Computer Science and Applications,*
*Vivekanandha College of Arts and Sciences for Women (Autonomous)*
*Elayampalayam, Tiruchengode, Namakkal(DT)-637205*

**Abstract: -** Efficient file query is important to the overall performance of peer-to-peer (P2P) file sharing systems. Clustering peers by their common interests can significantly enhance the efficiency of file query. Clustering peers by their physical proximity can also improve file query performance. However, few current works are able to cluster peers based on both peer interest and physical proximity. Although structured P2Ps provide higher file query efficiency than unstructured P2Ps, it is difficult to realize it due to their strictly defined topologies. In this work, we introduce a Proximity-Aware and Interest-clustered P2P file sharing System (PAIS) based on a structured P2P, which forms physically-close nodes into a cluster and further groups physically-close and common-interest nodes into a sub-cluster based on a hierarchical topology. PAIS uses an intelligent file replication algorithm to further enhance file query efficiency.

creates replicas of files that are frequently requested by a group of physically close nodes in their location. Moreover, PAIS enhances the intra-sub-cluster file searching through several approaches. First, it further classifies the interest of a sub-cluster to a number of sub-interests, and clusters common-sub-interest nodes into a group for file sharing. Second, PAIS builds an overlay for each group that connects lower capacity nodes to higher capacity nodes for distributed file querying while avoiding node overload. Third, to reduce file searching delay, PAIS uses proactive file information collection so that a file requester can know if its requested file is in its nearby nodes. Fourth, to reduce the overhead of the file information collection, PAIS uses bloom filter based file information collection and corresponding distributed file searching. Fifth, to improve the file sharing efficiency, PAIS ranks the bloom filter results in order. Sixth, considering that a recently visited file tends to be visited again, the bloom filter based approach is enhanced by only checking the newly added bloom filter information to reduce file searching delay. Trace-driven experimental results from the real-world PlanetLab testbed demonstrate that PAIS dramatically reduces overhead and enhances the efficiency of file sharing with and without churn.

**Keywords:** P2P Netwroks, file sharing system, Bloom filter.

## I. INTRODUCTION

OVER the past fewyears, the immense popularity of the Internet has produced a significant stimulus to P2P file sharing systems. For example, BitTorrent constitutes roughly 35 percent of all traffic on the Internet.

 

There are two classes of P2P systems: unstructured and structured. Unstructured P2P networks such as Gnutella and Freenet do not assign responsibility for data to specific nodes. Nodes join and leave the network according to some loose rules. Currently, unstructured P2P networks' file query method is based on either flooding where the query is propagated to all the node's neighbors, or random- walkers where the query is forwarded to randomly chosen neighbors until the file is found.

However, flooding and random walkers cannot guarantee data location. Structured P2P networks , i.e., Distributed Hash Tables (DHTs), can overcome the drawbacks with their features of higher efficiency, scalability, and deterministic data location. They have strictly controlled topologies, and their data placement and lookup algorithms are precisely defined based on a DHT data structure and consistent hashing function. The node responsible for a key can always be found even if the system is in a continuous state of change. Most of the DHTs require $O(\log n)$ hops per lookup request with $O(\log n)$ neighbors per node, where n is the number of nodes in the system. A key criterion to judge a P2P file sharing system is its file location efficiency. To improve this efficiency, numerous methods have been proposed. One method uses a super peer topology, which consists of super nodes with fast connections and regular nodes with slower connections. A super node connects with other super nodes and some regular nodes, and a regular node connects with a super node. In this super-peer topology, the nodes at the center of the network are faster and therefore produce a more reliable and stable backbone.

This allows more messages to be routed than a slower backbone and, therefore, allows greater scalability. Super-peer networks occupy the middle-ground between centralized and entirely symmetric P2P networks, and have the potential to combine the benefits of both centralized and distributed searches. Another class of methods to improve file location efficiency is through a proximity-aware structure. A logical proximity abstraction derived from a P2P system does not necessarily match the physical proximity information in reality. The shortest path according to the routing protocol (i.e., the least hop count routing) is not necessarily the shortest physical path. This mismatch becomes a big obstacle for the deployment and performance optimization

of P2P file sharing systems. A P2P system should utilize proximity information to reduce file query overhead and improve its efficiency. In other words, allocating or replicating a file to a node that is physically closer to a requester can significantly help the requester to retrieve the file efficiently. Proximity-aware clustering can be used to group physically close peers to effectively improve efficiency.

The third class of methods to improve file location efficiency is to cluster nodes with similar interests, which reduce the file location latency. Although numerous proximity-based and interest-based super-peer topologies have been proposed with different features, few methods are able to cluster peers according to both proximity and interest. In addition, most of these methods are on unstructured P2P systems that have no strict policy for topology construction. They cannot be directly applied to general DHTs in spite of their higher file location efficiency.

## II. RELATED WORK

We discuss the related works most relevant to PAIS in three groups: super-peer topology, proximity-awareness, and

 

interest-based file sharing. Super-peer topology. FastTrack [10] and Morpheus [20] use super-peer topology. The super-peer network in [8] is for efficient and scalable file consistency maintenance in structured P2P systems. Our previous work built a super-peer network for load balancing [9]. Garbacki et al. [21] proposed a self-organizing super-peer network architecture that solves four issues in a fully decentralized manner: how client peers are related to super-peers, how super-peers locate files, how the load is balanced among the super-peers, and how the system deals with node failures.

Proximity-awareness. Techniques to exploit topology information in P2P overlay routing include geographic layout,

proximity routing, and proximity-neighbor selection. Geographic layout method maps the overlay's logical ID space

to the physical network so that neighboring nodes in the ID space are also close in the physical network. It is employed in topologically-aware CAN [11]. In the proximity routing method, the logical overlay is constructed without considering the underlying physical topology.

Interest-base file sharing. One category of interest-base file sharing networks is called schema based networks. They use explicit schemas to describe peers' contents based on semantic description and allow the aggregation and integration of data from distributed data sources. Hang and Sia proposed a method for clustering peers that share similar properties together and a new intelligent query routing strategy.

Liu et al. proposed online storage systems with peer assistance. The works in employ the Bloom filter technique for file searching. Despite the efforts devoted to efficient file location in P2P systems, there are few works that combine the super-peer topology with both interest and proximity based clustering methods. In addition, it is difficult to realize in DHTs due to their strictly defined topology and data allocation policy. This paper describes how PAIS tackles the challenge by taking advantage of the hierarchical structure of a DHT.

## III. PROBLEM STATEMENT

### 3.1 Exiting Model

- A key criterion to judge a P2P file sharing system is its file location efficiency. To improve this efficiency, numerous methods have been proposed. One method uses a super peer topology which consists of supernodes with fast connections and regular nodes with slower connections. A supernode connects with other supernodes and some regular nodes, and a regular node connects with a supernode.
- In this super-peer topology, the nodes at the center of the network are faster and therefore produce a more reliable and stable backbone. This allows more messages to be routed than a slower backbone and, therefore, allows greater scalability. Super-peer networks occupy the middle-ground between centralized and entirely symmetric P2P networks, and have the potential to combine the benefits of both centralized and distributed searches.
- Another class of methods to improve file location efficiency is through a proximity-aware structure.

 

- The third class of methods to improve file location efficiency is to cluster nodes with similar interests which reduce the file location latency.

### 3.1.1 Disadvantages Of Existing System

- Although numerous proximity-based and interest-based super-peer topologies have been proposed with different features, few methods are able to cluster peers according to both proximity and interest.
- In addition, most of these methods are on unstructured P2P systems that have no strict policy for topology construction.
- They cannot be directly applied to general DHTs in spite of their higher file location efficiency.

### 3.2 Proposed System

- This paper presents a proximity-aware and interest-clustered P2P file sharing System (PAIS) on a structured P2P system. It forms physically-close nodes into a cluster and further groups physically-close and common-interest nodes into a sub-cluster. It also places files with the same interests together and make them accessible through the DHT Lookup() routing function. More importantly, it keeps all advantages of DHTs over unstructured P2Ps. Relying on DHT lookup policy rather than broadcasting, the PAIS construction consumes much less cost in mapping nodes to clusters and mapping clusters to interest sub-clusters. PAIS uses an intelligent file replication algorithm to further enhance file lookup efficiency.
- It creates replicas of files that are frequently requested by a group of physically close nodes in their location. Moreover, PAIS enhances the intra sub-cluster file searching through several approaches
- First, it further classifies the interest of a sub-cluster to a number of sub-interests, and clusters common-sub-interest nodes into a group for file sharing.
- Second, PAIS builds an overlay for each group that connects lower capacity nodes to higher capacity nodes for distributed file querying while avoiding node overload.
- Third, to reduce file searching delay, PAIS uses proactive file information collection so that a file requester can know if its requested file is in its nearby nodes.
- Fourth, to reduce the overhead of the file information collection, PAIS uses bloom filter based file information collection and corresponding distributed file searching.
- Fifth, to improve the file sharing efficiency, PAIS ranks the bloom filter results in order. Sixth, considering that a recently visited file tends to be visited again, the bloom filter based approach is enhanced by only checking the newly added bloom filter information to reduce file searching delay.

### 3.2.1 Advantages Of Proposed System

- The techniques proposed in this paper can benefit many current applications such as content delivery networks, P2P video-on-demand systems, and data sharing in online social networks.
- We introduce the detailed design of PAIS. It is suitable for a file sharing system where files can be classified to a number of interests and each interest can be classified to a number of sub-interests.
- It groups peers based on both interest and proximity by taking advantage of a hierarchical structure of a structured P2P.
- PAIS uses an intelligent file replication algorithm that replicates a file frequently requested by physically close nodes near their physical location to enhance the file lookup efficiency.

 

- PAIS enhances the file searching efficiency among the proximity-close and common interest nodes through a number of approaches.

## IV. OVERVIEW

### 4.1 PAIS: A proximity-aware interest-clustered p2p file sharing system.

In our previous work], we studied a BitTorrent user activity trace to analyze the user file sharing behaviors. We found that long distance file retrieval does exist. Thus, we can cluster physically close nodes into a cluster to enhance file sharing efficiency. Also, peers tend to visit files in a few interests. Thus, we can further cluster nodes that share an interest into a sub-cluster. Finally, popular files in each interest are shared among peers that are globally distributed.
Thus, we can use file replication between locations for popular files, and use system-wide file searching for unpopular files. We introduce the detailed design of PAIS below. It is suitable for a file sharing system where files can be classified to a number of interests and each interest can be classified to a number of sub-interests.

### 4.2 PAIS Structure

PAIS is developed based on the Cycloid structured P2P network. Cycloid is a lookup efficient, constant-degree overlay with $n=d.\ 2d$ nodes, where d is its dimension. It achieves a time complexity of $O(d)$ per lookup request by using $O(1)$ neighbors per node. Each Cycloid node is represented by a pair of indices $(k,a_{d-1}a_{d-2}....a_0)$ where k is a cyclic index and $(a_{d-1}a_{d-2}....a_0)$ is a cubical index. The cyclic index is an integer ranging from 0 to d - 1, and the cubical index is a binary number between 0 and 2d - 1. The nodes with the same cubical index are ordered by their cyclic index mod d on a small cycle, which we call a cluster.

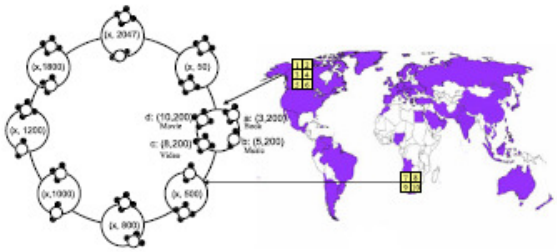

*Figure 1 : PAIS Struture*

### 4.3 PAIS Construction and Maintenance

Node proximity representation. A landmarking method can be used to represent node closeness on the network by indices used in. Landmark clustering has been widely adopted to generate proximity information. It is based on the intuition that nodes close to each other are likely to have similar distances to a few selected landmark nodes. We assume there are m landmark nodes that are randomly scattered in the Internet.

## V. EXPERIMENTAL RESULTS

We implemented a prototype of PAIS on PlanetLa , a real-world distributed testbed, to measure the performance of PAIS in comparison with other P2P file sharing systems. We set the experiment environment according to the study results of a BitTorrent trace. We randomly selected 350 PlanetLab nodes all over the world. Among these nodes, we randomly selected 30 nodes as landmark nodes to

calculate the Hilbert numbers of nodes. We clustered all nodes into 169 different locations according to the closeness of their Hilbert numbers.

We used the 56,076 files in the BitTorrent trace. The number of interests in the system was set to 20, so we also set the dimension of the Cycloid DHT to 20. We simulated 100,000 peers by default in the experiments. Each peer was randomly assigned to a location cluster among all 169 clusters, and further randomly assigned to a Planet- Lab node within this location. According to, a peer's requests mainly focus on around 20 percent of all of its interests. Thus, we randomly selected four interests (20 percent of total 20 interests) for each peer as its interests.

The files are randomly assigned to a sub-cluster with the files' interest over the total 160 locations, and then randomly assigned to nodes in the sub-cluster. Eighty percent of all queries of a requester target on files with owners within the same location, among which 70 percent of its queries are in the interests of the requester. According to [48], 80 percent of all requests from a peer focus on its interests, and each of other requests is in a randomly selected interest outside of its interests. A request in an interest means a request for a randomly selected file in this interest. We also let each file have a copy in another peer in a different location in order to test the proximity-aware file searching performance.

## VI. CONCLUSION

In recent years, to enhance file location efficiency in P2P systems, interest-clustered super-peer networks and proximity- clustered super-peer networks have been proposed. Although both strategies improve the performance of P2P systems, few works cluster peers based on both peer interest and physical proximity simultaneously. Moreover, it is harder to realize it in structured P2P systems due to their strictly defined topologies, although they have high efficiency of file location than unstructured P2Ps.

In this paper, we introduce a proximity-aware and interest-clustered P2P file sharing system based on a structured P2P. It groups peers based on both interest and proximity by taking advantage of a hierarchical structure of a structured P2P. PAIS uses an intelligent file replication algorithm that replicates a file frequently requested by physically close nodes near their physical location to enhance the file lookup efficiency. Finally, PAIS enhances the file searching efficiency among the proximity-close and commoninterest nodes through a number of approaches. The trace-driven experimental results on PlanetLab demonstrate the efficiency of PAIS in comparison with other P2P file sharing systems. It dramatically reduces the overhead and yields significant improvements in file location efficiency even in node dynamism. Also, the experimental results show the effectiveness of the approaches for improving file searching efficiency among the proximityclose and common-interest nodes.

## REFERENCE

[1] BitTorrent. (2013) [Online]. Available: http://www.bittorrent.com/
[2] Gnutella home page. (2003) [Online]. Available: http://www.gnutella.com
[3] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong, "Freenet: A distributedanonymous information storage and retrieval system," inProc. Int. Workshop Des. Issues Anonymity Unobservability, 2001,pp. 46–66.

[4] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek,F. Dabek, and H. Balakrishnan, "Chord: A scalable peer-topeerlookup protocol for internet applications," IEEE/ACM Trans.
Netw., vol. 11, no. 1, pp. 17–32, Feb. 2003.

[5] A. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems," in Proc. IFIP/ACM Int. Conf. Distrib. Syst. Platforms Heidelberg, 2001, pp. 329–350.
[6] B. Y. Zhao, L. Huang, J. Stribling, S. C. Rhea, A. D. Joseph, and J. Kubiatowicz, "Tapestry: A resilient global-scale overlay for service deployment," IEEE J. Sel. Areas Commun., vol. 22, no. 1, pp. 41–53, 2004.
[7] H. Shen, C. Xu, and G. Chen, "Cycloid: A scalable constant-degree P2P overlay network," Perform. Eval., vol. 63, pp. 195–216, 2006.
[8] Z. Li, G. Xie, and Z. Li, "Efficient and scalable consistency maintenance for heterogeneous peer-to-peer systems," IEEE Trans. Parallel Distrib. Syst., vol. 19, no. 12, pp. 1695–1708, Dec. 2008.
[9] H. Shen and C.-Z. Xu, "Hash-based proximity clustering for efficient load balancing in heterogeneous DHT networks," J. Parallel Distrib. Comput., vol. 68, pp. 686–702, 2008.

 