Efficient File Querying In Locality Aware Interest Cluster Peer To Peer File Sharing System

S.Gowsika' M.E.,

Abstract: In peer to peer networking, in order to improve the overall performance of the file sharing system the competent file querying is essential, for this a method called Locality aware interest cluster p2p file sharing system (LAIC) is introduced. In this work, based on the locality the physically close nodes are clustered and based on the interest sub cluster is formed. Locality aware interest clustering enhances the process through various steps. First, the proximity nodes are clustered and the interest nodes are classified into sub-cluster as a group. Second, it constructs an overlay to diminish node overload. Third, it reduces the file sharing setback by using proactive file information collection. Fourth, the overhead of the file information collection is reduced by using bloom filter based information. Fifth, the Cluster leader is elected by using JSD method. In this work, file querying and sub-interest file querying mechanism is proposed to increase the efficiency in interest cluster. To further enhance the file query efficiency, a cost effective file replication algorithm is introduced In-order to reduce the cost as well as to improve the query efficiency.

Keywords: P2P; LAIC; Bloom filter; File replication; cluster.

I. INTRODUCTION

P2P stands for peer to peer network. The term peer means that the computer system or nodes. The vast esteem of the internet has shaped a major motivation to P2P file sharing systems. A P2P network is a type of decentralized and distributed architecture. In this type of network, information is Shared between multiple peers that access the resources directly with the help of other network with any servers. P2P networking utilize the Internet for client-based computing tasks and in modern world personal computers have speedy processor, huge memory, and a bulky hard disk so that they can perform the common computing tasks like e - mail and web browsing. A server computer usually has large number of resources and must respond to the resources for the data transaction between many client computers. Client computers kick off requests for resources or data from server computer. Many major works has been done for increasing the performance and for the efficient information sharing in peer to peer network. Peer to Peer network is classified into two types, Unstructured Peer to peer network and structured peer to peer network and the main principle involved in P2P network are sharing resource, self organization and decentralization.

A. Unstructured Peer to Peer network:

In this network does not inflict on a particular structure by their design on the overlay network. In unstructured P2P the file querying process is done either by flooding or by the random walker's. In this method the query is propagated to all the neighbour nodes which is chosen at random, the query forwarding will continue until the file found during this searching process. Although it processes the query well, it does not guarantee the data location.

B. Structured Peer to Peer network:

Distributed Hash Table (DHT) is used, in which the neighbour relationship between the node and the data locations are strictly defined. Distributed hash table is a type of decentralized and distributed system that provides lookup service. As in the hash table the lookup services are similar. (value, key) pair is stored in DHT in which the values associated with a given key can be efficiently retrieved by the participating node. Even though the system is in a state of change, the node which is in charge for a key can be found.

Among structured system, the DHT is implemented using the different data structure. The DHT structure is classified into abstract key space and random unique key (identifier). The abstract key space consists of large inter values (range from 0 to 2^{128}-1). From this random unique (identifier) key space is assigned to each participant. All node maintains a small routing table consisting of its neighboring peer. These routing information are linked together to form a overlay network. The routing procedure will traverse among the nodes based on the routing information to reach the destination. This process is sometimes called as key based routing.

II. RELATED WORK

The work of LAIS is mostly related to the cluster based on the locality and some approaches that enhance the location efficiently. The super node network is mainly referred for their scalability, efficiency and file consistency maintenance in structured peer to peer network. D.H. Epema et all.[3] introduced an architecture called self organizing super peer network architecture (SOSPNET). It is usually constructed on the peak of unstructured topology with semantic association among the nodes and the files. There are two different peers involved in this architecture one is super peer and the weak peer. Here the information related to the content is stored in the super peer. Weak peer will sort the super peer. The sorting is done according to the number of positive response to the quires and connection is done to the super peers which suit the majority of their request. It solves some of the problems in a fully decentralized approach: how super peer situate the files, how client peers are linked to each other, how the load balancing is maintained among the super peers and how the system deal with the peer failure. There are some techniques to develop each topology information in p2p overlay steering includes proximity awareness i.e. neighbor selection, routing and geographic layout. Antony Rowstron et all.[1] proposed design and development of pastry, which is similar to that of chord. PASTRY assigns 128 bits node id to each node in the system and every peer is in charge for handling and routing requests for numeric keys to node with closest node id($B=2^b$). In pastry each and every node maintains three structures such as routing table, a neighborhood set, and a leaf set. Routing table consists of ($\log_B N$) rows and B-1 columns (B is the configuration

parameter with typical value 4). The row in each cell has routing information eg: the IP address of all nodes whose identifier has same n first digit as latest node. The locality n set contains the node that are close in its proximity to its current node (2*B). The leaf set contains |L|/2 nodes which is numerically closest and greater than its current node. Using the node Id which is in the desired portion of the id space the entries in the routing table are selected. Fanbin Meng [2] projected a hierarchical clustering Peer to Peer network model. It is based on user interest to improve search efficiency by using the topological algorithm. This algorithm processes the received query and will send back the reply with decreased copious arbitrary penetrating process and gains the appropriate supply quicker than usual searching algorithm. Haiying Shen. [5] Proposed a hash based proximity clustering, which is based upon the consistent hashing function. This method balances the work load among each and every node in the network. Here the clustering occurs in the physical network and in virtual network. The cluster in the physical network is termed as pcluster and virtual network is termed as vcluster. In pcluster the regular node are associated to their physically nearest super node and occasionally reports the load information to the super node whereas in vcluster regular nodes are connected to the rationally close super node as in novel DHT network. Although these clusters are self organized, it still need for least amount preservation as in DHT network. Haiying Shen et all. [4] Proposed nearness aware interest clustering concept in structured peer to peer network. This method is urbanized based on the cycloid peer to peer network. Cycloid is a lookup competent stable degree superimposed network with $n=d.2^d$ (d is the dimension)of the peer. PAIS forms the cluster based upon the node proximity and interest by using the consistent hashing function.

III. EXISTING SYSTEM

In exsisting method in order to get better file querying performance and to improve the competency of file querying the peers in the p2p network are clustered. Here clustering is done by proximity and sub clustering is by grouping the interest file together. The interest node within the cluster is constructed by using SHA-1 hash hashing function. Thus a super peer topology is formed (each super peers are connected to form super peer topology).After clustering the files in the clusters are accessed by using DHT lookup () routing function The cluster in PAIS acts as super peer network, the server in a sub group acts as a central server to its compartment of clients by maintaining a directory of files in the clients. When a client request for a file by sending query to the server and receive the location result if the file is found. Servers are also associated with each other as nodes in cycloid. Server forward message over superimpose and surrender and answer queries by representing their clients and themselves. Each cluster consists of super nodes which represents the interest file. In existing method super node only stores id of all the files in a peer. While searching for a file or uploading a file the process becomes tedious irrespective of the id. To reduce the time delay during this process the cluster leader can be elected by computing the total JSD measure value in the intra cluster. The cluster leader stores the path of all the sub interest files. So the file searching efficiency can be improved. To further enhance the file location efficiency the existing method uses intelligent file replication method. When the requested regularity of a file from a cluster of nodes with its Id exceed the threshold value the files will be replicated. The major drawback in PAIS during replication is that, requests that are sent out from nodes in a site create more replicas for dissimilar file. More replicas will reduce the performance of the peer and there is a chance of peer crashes, which significantly increases the cost

IV.PROPOSED METHODOLOGY

In proposed work each of the cluster is formed based on the locality (physically close nodes as a cluster) and based on their interest .This methodology gives a new concept of resource for the interest clustering and also for file replication process.

- The interest nodes are classified into sub-cluster as a group.
- Overlay network is constructed to reduce the node overload.
- By using proactive file information collection the file sharing delay is reduced.
- The overhead of the file information is reduced using bloom filter information.
- Cluster leader is elected using JSD method to fast access
- To further reduce the delay the cost effective file replication algorithm is used, where the underutilized replicas are removed

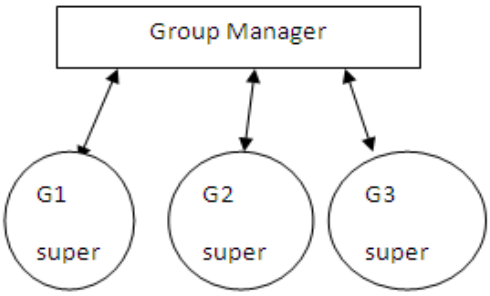

***Fig.1**.a Group based Arbitration*

MODULES

A. NODE LOCALITY AND INTEREST ILLUSTRATION:

Since we are using DHT structured protocol the cycloid overlay is used. A methodology known as Land marking is used to produce proximity information based on locality. It is based on the instinct that peers just about each different are prone to have identical distances to a couple chosen landmark nodes. Two bodily close nodes will have equivalent vectors. The closeness of two nodes' Hs indicates their substantial closeness on the internet. In land mark method space filling curve called Hilbert curve is used. It maps the m-dimensional landmark vectors to actual numbers, so the closeness relationship among the nodes is preserved. We identify this variety the Hilbert number of the node denoted by means of H. For determining the file interest the SHA-1 hash function can be used

B. FILE QUERY MECHANISM

To inquest a file or to upload a file the server maintains the directory of all files in its cluster leader. For every request,

the server searches for a file in the intra-cluster and inter-cluster. Initially intra-cluster search is performed. In this search the server check for the requested file with key(id) . If the corresponding key is found, then the node sends the locality of the file to its requestor. In the other case it performs the inter-cluster probing. During the intra and inter cluster probing in order to enhance the efficiency of the file searching the DHT lookup() is prepared among the nodes in peer to peer network.

C. SUB-INTEREST FILE QUERYING MECHANISM

In proposed Locality-Aware Interest Clustering method especially for a video file a narrative JSD-based hunt method is used. In Jenshon Shannon divergence method the query processing and the forwarding is done based on the JSD value. The rage of JSD should stuck between the requestor interest and the requested video interest.

Calculating JSD value: To calculate the JSD the below given formula is used.

$JSD(P_i(v),P_j(v))=1/2(D_{KL}(P_i(v)||(P_i(v)+P_j(v)/2))+D_{KL}(P_j(v)||(P_i(v)+P_j(v)/2)))$ (1)

V-words in the vocabulary of all peers

$P_i(v)$-word frequency histogram in i peer

v€V- word in vocabulary

$p_i(v)$- % of word in $P_i(v)$

D_{KL}-Divergence between peer i and j

D_{KL} computed using the below given formula

$D_{KL=}(P_i(v)||P_j(v))=\sum_v p_i(v)\log P_i(v)/P_j(v)$ (2)

Peer interest set has a number of keywords with weight linked with it and this represents the importance or the frequency of a keyword from the node's view.

Leader election: The election process run as part during this technique inside the cluster, if joining of node and leaving of node exceeds a fixed number of times the election process exceeds, the cluster with the least total distance is elected as a leader for that particular process. The distance is computed by every peer total JSD remoteness within cluster.

C.COST EFFECTIVE FILE REPLICATION

To further improve the efficiency of file searching process with less delay a concept of file duplication process is introduced. This process deals with a distributed cost effective file replication algorithm that can roughly comprehend the file replication rule in a distributed manner. This algorithm will replicate the file (frequently requested files) and adjustments among the nodes are done adaptively. The file replication depends upon its query rate and the threshold value (constant parameter). Query rate is represented by q_f , is the number of request per unit time. A threshold value Tq - query rate of a file, which is denoted as Tq=α * avgq(α is a constant parameter,) α >2 and Avg_q - average query rate and it is computed using the formula

$$Avg_q=\sum_{j=1}^{n}(qfj)\ n/$$

In this decentralized approach the replicas can be created and also be removed. This process i.e. the creation and removal depends upon the calculation of a node's query rate. In node, if the request rate is less, the underutilized(replicas which are not used)will be removed from that particular node. In case, the request rate of a peer q_r> δ Tq(δ is the under loaded factor i.e. δ>1) the replicas are noted even if this situation occurs once. For a exacting time interval, it will be detached permanently. The requestor of a file will calculate the request rate if it is greater than t_q the file is queried including the replication request.

Calculation of cost for replication: The cost for file replication can be calculated by considering each upload operation of a file.

$$\text{Cost (Update)} =(x+a)^n=\sum_{i=1}^{u}\text{size of (file)}*\text{dst}$$

U- Number of files for update operation
Dst- files deliver at dst hop
Size -It can be set upto1000 bytes
File query Size –It can be set up to 27 bytes.

V. PERFORMANCE ANALYSIS

In this paper based on the locality the proximity and the interest nodes are taken into account for clustering. The landmark methodology is implemented for calculating Hilbert number of nodes. The user interest of a file is calculated based on the priority of a file and using JSD value the cluster leader is elected where as all the indices of a files are stored. Here the number of file querying forwarded to cluster leader and the response from cluster leader is measured. To further improve the performance the file replication is done. The replicas of a files are stored at the cluster leader, where as each underutilized replicas are removed then and there based on the query rate of a file. So compared to previous work this approach is able decrease the cost especially the locality ignorant.

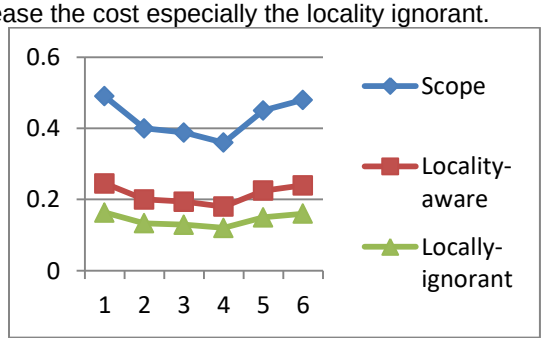

***Fig. 2** . performance analysis graph*

VI. CONCLUSION

In this paper enrichment of file location efficiency, clustered locality p2p and interest clustering has been proposed. Although these approaches improve the recital, but the physical nearest cluster and peer interest the competence can be still faster. In this cluster based on location peer to peer system and interest cluster peer to peer file sharing system in structured network would swiftly increase the performance rate of identifying file location. LAIC method uses a cost effective file replication to enhance physical locality of frequently accessed nodes for improving competence. Although the cost effective replication improves the efficiency, Careful selection of the threshold value is necessary.

REFERENCES

[1] Antony Rowstron1 and Peter Druschel "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems"

[2] Fanbin Meng, Lei Ding, Sheng Peng and Guangxue Yue “A P2P Network Model Based on Hierarchical Interest Clustering Algorithm” in journal of software, vol. 8, no. 5, may 2013
[3] Garbacki, P., D.H. Epema and M. Van Steen,Self-Organizing Super-Peer Networks.
[4] Haiying Shen, Senior Member, IEEE, Guoxin Liu, Student Member, IEEE and Lee Ward “A Proximity-Aware Interest-Clustered P2P File Sharing System” in ieee transactions on parallel and distributed systems, vol. 26, no. 6, june 2015
[5] S.Saravanan, Arivarasan.”An efficient ranked keyword search for effective utilization of outsourced cloud data” Journal of Global Research in Computer Science, 2035, Vol4(4), pp:8-12
[6] Haiying Shen and Cheng-Zhong Xu, “Hash-based Proximity Clustering for Load Balancing in Heterogeneous DHT Networks” in journal of parallel and distributed computing, vol.68 Issue 5, may 2008