International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

Self-organized file sharing by node feature in clustered P2P system

[1] Anju S Kumar [2] Ratheesh S
[1] PG Scholar, [2] Assistant Professor
College of Engineering, Perumon, Kerala, India
[1] anjuskumar16@gmail.com, [2] ratheeshs2007@gmail.com

Abstract- **The overall performance of peer-to-peer (P2P) file sharing lies on the efficiency of file query. To enhance the efficiency of file query in a structured peer-to-peer system, clustering technique can be used. Clustering peers by their common interests and by their physical proximity can improve file query performance. In the clustering technique the physically-close nodes are formed into a cluster and further physically-close and common-interest nodes are grouped into a sub-cluster based on a hierarchical topology. In the search mechanism, the file query will move to the nearest proximity cluster and to the corresponding interest cluster within that proximity cluster. If all nodes within this cluster can respond to the query, there is a need for a method to choose appropriate node to share the file. In this paper, we propose a method, called the statistical feature matrix form (SMF), which improves the searching mechanism in the structured Peer–to-Peer system by selecting neighbors according to their capabilities. SMF measures the number of shared files, the content quality, the query service and the transmission distance between neighbor nodes. Based on these measurements, appropriate nodes can be selected by finding the rank of each nodes in the cluster, thereby reducing the traffic overhead significantly and enhance the file sharing efficiency in the structured P2P system.**

Index Terms— **Clustering, Characteristics matrix, Rank matrix , Weight matrix.**

I. INTRODUCTION

The performance optimization and efficiency improvement of content sharing in peer-to-peer (P2P) networks requires a significant amount of work. There are two classes of P2P systems: unstructured and structured. Unstructured peer-to-peer networks do not impose a particular structure on network by design, they are formed by nodes that randomly form connections with each other, where file query method is based on either flooding or random- walkers. File location efficiency is the key criterion to judge a P2P file sharing system. In the Unstructured P2P system the data location cannot be guaranteed. In the structured P2P system the Distributed Hash Table(DHT) technique can provide higher efficiency and deterministic data location.

Super-peer network topology system is one method to improve file querying and file location efficiency. This super-peer network topology consists of higher connectivity super nodes and regular nodes with low connectivity. A super node connects with other super nodes and some regular nodes. The regular node connects with a super node. Clustering of nodes in the network is another method to improve file location efficiency in structured P2P system. File replication technology is also widely used to reduce hot spots and thereby improving the file query efficiency.
Clustering nodes based on the interest and proximity of the node is the efficient technique clustering in the Peer-to-Peer system. In this, physically close nodes in the network are formed into a cluster and further group physically-close and common-interest nodes into a sub-cluster [1]. Files having the same interests are placing together as sub cluster in each cluster and these files are made accessible through the DHT Lookup () routing function. Thereby the file querying in the network will be more efficient.

In the file querying technique the query will move to the nearest proximity cluster and within this proximity the query will again be forwarded to corresponding sub cluster for the file location. In a sub-cluster there may a number of files which will be capable of sharing the required file. In such a case we do the DHT Look up method for finding the nodes which can share the file in a minimum hopes. For the nodes with same hope we have to choose appropriate node for the file sharing[9]. This paper proposed a method for choosing a node from a number of nodes which are at same hope from the requesting node. A method called the Statistical feature matrix form (SMF) improves the querying mechanism by selecting nodes according to their capabilities. In this nodes are analyzing in terms of the following characteristics of the nodes: *Processing Ability*

(*PA*), *Effective Sharing* (*ES*). SMF improves the file querying method and thereby enhance the file sharing efficiency.

II. LITERATURE REVIEW

The most relevant works related to this efficient file sharing by node clustering are:

- Super-peer topology
- Proximity-awareness
- Interest-based clustering
- Consistency maintenance

Consistency maintenance and load balancing can be achieved through the superpeer topology, which exploit the heterogeneity of nodes in a peer-to-peer (P2P) network by assigning additional responsibilities to high capacity nodes called super-peers. Weak peers submit queries to their super-peers and receive results from them.

Proximity Awareness is the first step in the clustering technique in the P2P file sharing system[2]. The physically close nodes are grouped together and within this physically close nodes only further classifications of the nodes can be done . Node closeness can be represented by a method called Landmark method.

For the interest identification a method of signature calculation is used[4]. Each peer is having a collection of data and these data collection is prepossessed and calculate a signature value SIG to characterize their data properties. Thus each signature value represents the interest associated with each node.

Consistency maintenance mechanism is necessary for the peer-to- peer (P2P) system due to their frequent data updates. For each replica group an overlay network is established with two layers. The upper layer is Distributed Hash Table (DHT) based and consists of powerful and stable replica nodes called Chord Replica Nodes (CRN).The lower or the second layer consists of the Ordinary Replica Nodes (ORN)[5]. On specific update, the update massage passes through the tree and every replica nodes receive the update message.

III. PROPOSED SYSTEM

In the clustered file sharing System based on a structured P2P network, the physically-close nodes are formed into a cluster and further physically-close and common-interest nodes are grouped into a sub-cluster based on a hierarchical topology. This clustered system uses an intelligent file replication method to replicate a files that are frequently requesting by physically close nodes near their physical location to enhance the file lookup efficiency and thereby enhance the overall performance of file sharing in the P2P system[6]. In the search mechanism, the file query will move to the nearest proximity cluster and to the corresponding interest cluster within that proximity cluster. If all nodes within this cluster have the same file for the search, there is a need for a method to choose appropriate node to share the file[7].In this case we have to consider some of the capabilities of all that nodes and find the most capable node among them. We propose a method that statistically analyzes query messages in terms of the following two characteristics: *Processing Ability* (*PA*), *Effective Sharing* (*ES*). The *PA* of peers is analyzed to determine which peers leech the most resources by processing the query. The *ES* refers to the number of files that a peer shares, and can be used to classify a peer's sharing capability [9]. It has been shown that, in a network, very few peers share a large number files, so that the quality of the files influences the sharing capability[9].

The SMF of query peer *u* is comprised of two matrixes: the *left-hand matrix* and the *right-hand matrix*. The *left-hand matrix*, called the *characteristics matrix* (*CM*), is an n x 2 matrix, where *n* is the number of nodes to which query is passing. The right-hand matrix, called the *weight matrix* (*WM*), is a 2x1 matrix in which each peer can set the proper weights according to the derivation degree of each characteristics. Finally, each query peer computes a *ranking matrix* (*RM*), which is an *n*x1 matrix obtained by the matrix multiplication *CM* x *WM* and then select the nodes with the top-*k* rank to send query messages.

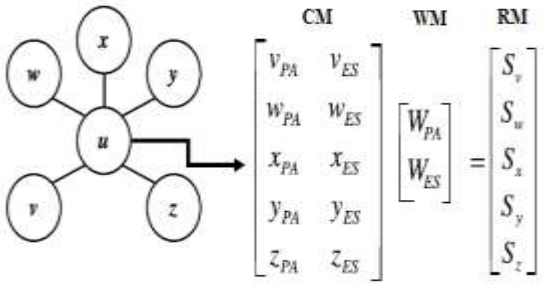

Fig1: An example of the SMF for a query peer u.

A. Characteristics matrix generation

The *Processing Ability* (*PA)* score is computed in terms of the *peers' query frequency* (QF)and *response frequency(RF)* .The *effective sharing* (ES), which is used to determine the number of files shared among peers in a P2P network. The ES is comprised of two sub-features: the *sharing count* (*SC*) and the *quality of sharing* (*QS*).

B. Query Frequency (QF)

Let *N*(*u*) be the neighbors of a query peer *u*; that is, *N*(*u*) are peers that are one hop away from *u*. In addition, let *Q*(*v*) be the number of queries sent by *v*. Each query peer *u* computes *TQ*1(*u*), which is the total number of queries (TQ) sent from the peers that are one hop away from *u*. Formally,

$$TQ_1(u) = \sum_{v \epsilon N(u)} NQ(v).$$

The Query-Score (QS) of a neighbor v of u is defined as

$$QS(u, v) = TQ_1(u) - NQ(v).$$

each query peer *u* computes *TQS*1(*u*) (resp. *TQS*2(*u*)), which is the sum of the query-scores (*TQS*) of all peers that are one (resp. two) hop(s) away from *u*:

$$TQS_1(u) = \sum_{v \epsilon N(u)} QS(u, v)$$

and

$$TQS_2(u) = \sum_{v \epsilon N(u)} TQS_1(u)$$

Then, the *query frequency of a neighbor v of u* is defined as QF(v)
where *h1* and *h2* are two parameters used to adjust the influence of peers that are one hop away and two hops away from *u* respectively

$$QF(u, v) = h1 * \frac{QS(u, v)}{TQS_1(u)} + h2 * \frac{TQS_1(v)}{TQS_2(u)}$$

IV. RESPONSE FREQUENCY (RF)

The term "response frequency" refers to a peer's ability to respond to queries.
Each peer *u* computes *TR*1(*u*) (resp. *TR*2(*u*)), which is the sum of the response times (TR) of peers that are one (resp. two) hop(s) away from *u*. Formally,

$$TR_1(u) = \sum_{v \epsilon N(u)} NR(v)$$

where *NR*(*v*) is the number of responses sent by node *v*, and

$$TR_2(u) = \sum_{v \epsilon N(u)} TR_1(v).$$

The *response frequency of a neighbor v of u*, denoted by *RF*(*u*; *v*), is computed as follows:

$$RF(u, v) = h1 * \frac{NR(v)}{TR_1(u)} + h1 * \frac{TR_1(v)}{TR_2(u)}$$

where *h1* and *h2* are two parameters used to adjust the influence of peers that are one hop away and two hops away from *u* respectively.
We represent the *processing ability of a neighbor v of u*, denoted by *PA*(*u*; *v*), in terms of the query frequency and the response frequency as

$$PA(u, v) = QF(u, v) + RF(u, v).$$

V. SHARING COUNT (SC)

The number of files shared by the peers. If a peer shares a large number of files, it should have a higher probability of matching queries. Each query peer *u* computesT*SF*1(*u*) (resp. T*SF*2(*u*)) which is the total number of shared files (T*SF*) by peers that are one (resp. two) hop(s) away from *u*.

$$TSF_1(u) = \sum_{v \epsilon N(u)} NF(v)$$

And

$$TSF_2(u) = \sum_{v \epsilon N(u)} SF_1(v)$$

The *sharing count of a neighbor v of u* is defined as

$$SC(u, v) = h1 * \frac{NF(v)}{TSF_1(u)} + h2 * \frac{TSF_1(v)}{TSF_2(u)}.$$

QUALITY OF SHARING (QS)
It distinguish useful files from useless files.

$$TFH_1(u) = \sum_{v \epsilon N(u)} NFH(v)$$

and

$$TFH_2(u) = \sum_{v \epsilon N(u)} TFH_1(v).$$

The *quality of sharing* (*QS*) *of a neighbor v of u* is defined as QS(u, v)
We then represent the *effective sharing of a neighbor v of u*, denoted by *ES* (*u*, *v*)

$$QS(u, v) = h1 * \frac{NFH(v)}{TFH_1(u)} + h2 * \frac{TFH_1(v)}{TFH_2(u)}$$

$$ES(u, v) = SC(u, v) + QS(u, v).$$

Weight matrix generation

It is used to normalize the values of the feature matrix. we can set a higher value in the weight matrix to represent the diversity of the feature if the values of a feature are widely dispersed. we set a lower value in the weight matrix to reduce the influence of the feature if they are so close[9][10].

We adopt the mean and standard deviation techniques to achieve this goal.

VI. CONCLUSION

This paper, we introduce a clustered file sharing System based on a structured P2P network, in which physically-close nodes are formed into a cluster and further physically-close and common-interest nodes are grouped into a sub-cluster based on a hierarchical topology. In the search mechanism, the file query will move to the nearest proximity cluster and to the corresponding interest cluster within that proximity cluster. If all nodes within this cluster can respond to the query, there is a need for a method to choose appropriate node to share the file. In this paper, we propose a method, called the statistical feature matrix form (SMF), which improves the searching mechanism in the structured Peer–to-Peer system by selecting neighbors according to their capabilities. This clustered system uses an intelligent file replication method to replicate a files that are frequently requesting by physically close nodes near their physical location to enhance the file lookup efficiency and thereby enhance the overall performance of file sharing in the P2P system. This method of Statistical feature matrix form (SMF) improves the querying mechanism by selecting nodes according to their capabilities. In this nodes are analyzing in terms of the following characteristics of the nodes: *Processing Ability* (*PA*), *Effective Sharing* (*ES*). SMF improves the file querying method and thereby enhance the file sharing efficiency. SMF can reduce the traffic overhead significantly, achieve shorter query responded times, and maintain a high success rate .Based on the rank matrix the node with high score is chosen for the file sharing

REFERENCES

[1]. Haiying Shen, Senior Member, IEEE, Guoxin Liu, Student Member, IEEE and Lee Ward ”A Proximity-Aware Interest- Clustered P2P File Sharing System” IEEE transactions on parallel and distributed systems, vol. 26, no. 6, June 2015.

[2]. P. Garbacki, D. H. J. Epema, and M. V. Steen ”The design and evaluation of a self-organizing super-peer network” IEEE Trans. Comput., vol. 59, no. 3, pp. 317331, Mar. 2010.

[3]. . P. Garbacki, D. H. J. Epema, and M. van Steen ”Optimizing peer relationships in a super-peer network” in Proc. Int. Conf. Distrib. Comput. Syst., 2007, p. 31.

[4]. 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. 16951708, Dec. 2008.

[5]. 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. 686702, 2008.

[6]. C. Hang and K. C. Sia, ”Peer clustering and firework query mode” in Proc. Int. World Wide Web Conf., 2002.

[7]. G. Liu, H. Shen, and L. Ward, “An efficient and trustworthy P2P and social network integrated file sharing system,” Proc. P2P, 2012, pp. 203–213.

[8]. K. Elkhiyaoui, D. Kato, K. Kunieda, K. Yamada, and P. Michiardi,“A scalable interest-oriented peer-to-peer pub/sub network,” in Proc. 9th Int. Conf. Peer-to-Peer Comput., 2009, pp. 204–211.\

[9]. M. Yang and Y. Yang, “An efficient hybrid peer-to-peer system for distributed data sharing,” IEEE Trans. Comput., vol. 59, no. 9, pp. 1158–1171, Sep. 2010.

[10]. E. Adar and B. A. Huberman, ``Free riding on Gnutella," *First Monday*, vol. 5, nos. 10_12, pp. 1_22, Oct. 2000.

[11]. S. Bianchi, P. Felber, and M. G. Potop-Butucaru, ``Stabilizing distributed R-trees for peer-to-peer content routing," *IEEE Trans. Parallel Distrib.Syst.*, vol. 21, no. 8, pp. 1175_1187, Aug. 2010.