Node Reputation-Based Replica Replication Strategy in P2P Networks

Yongqiong Zhu
School of Art, Wuhan Business University, Wuhan, China
zyqzhuyongqiong@126.com

Abstract. Replica technology increases the reliability and availability of data by increasing the redundancy of data in the system. The existing replica replication algorithm only depends on the access traffic of the node when selecting the replication node, ignoring the non-cooperation of the node, resulting in a decrease in the utilization of the replica. This paper proposes an adaptive replica management policy PCA based on node reputation. PCA selects nodes with large query traffic and high reputation as replication nodes to ensure high query hit rate. Because PCA sets the copy according to the cooperation of the node, it avoids the waste of copy caused by setting the copy on the non-cooperative node, and guarantees high copy utilization.

1. Introduction
P2P networks are widely used in file sharing, instant messaging, distributed storage, and collaborative computing. Although there are many applications for P2P networks, these applications are based on its most basic function - file location. Since users' requests for files in P2P systems tend to be biased towards some of the most popular files, there are often some nodes in the network that carry popular files overwhelmed to form "hotspots".

The existence of a "hotspots" consumes a lot of bandwidth on the server, resulting in inefficient file sharing. An effective solution is to use file replication technology, that is, to create multiple replica of the same hotspot file in the network. Through copy replication technology, establishing certain data redundancy in the network can significantly improve the efficiency and availability of the network.

P2P networks are based on the assumption that nodes are willing to share resources. However, the actual situation is that the node's autonomous behavior makes the node's capability show great heterogeneity. There are a large number of nodes in the consumption network that only contribute resources from other nodes, but never contribute their own resources. Studies have shown that 70% of the nodes in the Gnutella network are free-rider nodes. The non-cooperation caused by this autonomy of nodes seriously affects the availability of p2p services. Although the current research proposes many solutions for P2P copy management, these programs lack analysis and attention to node non-cooperation. Since the non-cooperative node does not provide its own service, setting a copy on it will not improve the network performance, but will waste network resources, which makes the existing copy management strategy difficult to adapt to complex network environment, which is the main factor affecting P2P systems to provide high availability data services. Therefore, how to select the appropriate node for replica replication based on the cooperation of nodes is an important issue to be studied.
In recent years, research on node cooperation has been paid more and more attention by many scholars and put forward many evaluation models about the performance of node cooperation. In particular, the research found that P2P networks and social networks are similar, and the reputation mechanism in social networks is introduced into P2P networks, so that the reputation of nodes before the transaction is an effective means to improve the success rate of transactions. Based on this, this paper proposes a trusted copy replication policy PCA based on the reputation of the node. The novel feature of PCA is the ability to achieve high service satisfaction, high file query rates and high copy utilization with significant low overhead. Unlike previous replication algorithms, PCA chooses to query nodes with high traffic and high reputation as replica nodes to ensure high query hit rates. Because the PCA sets the copy according to the cooperation of the node, it avoids the waste of copy caused by setting the copy on the non-cooperative node, significantly reducing the number of useless copies and ensuring high copy utilization.

The structure of this paper is as follows. In the second part, the related work of the research and the objectives of this research are introduced. The third part introduces the system model of this paper, and gives the way to achieve it. Finally, the PCA algorithm is given. The fourth part is experimental simulation to prove our method. The fifth part is the conclusion.

2. Related work
In the current study of unstructured P2P replica replication, Cristina L. [1], Dongming Huang [2], Mohammadi [3] and Haiying Shen [4] and others have proposed a copy management strategy. According to the EAD algorithm proposed by Haiying Shen, EAD will use heuristic algorithms to select high access rate nodes that can minimize the search cost. However, the EAD algorithm does not consider the existence of uncooperative nodes. Under the reputation mechanism, a node often provides good services for everyone, and it can gain everyone's trust. The reputation of the node is high. On the contrary, if the node often does not provide services or provides unsatisfactory transactions, its reputation is the degree is very low, so that future nodes try to avoid these low-reputation nodes when selecting service nodes.

Regarding the research of node reputation, researchers at home and abroad have proposed a series of reputation models. Yuh-Jzer Joung [5] noted that free riding is a very common phenomenon in P2P networks. It establishes a global indexing service for files shared in the network, so it can avoid query messages from non-free riders when forwarding. Pass to free riders. Chithra Selvaraj, Sheila Anand [6] proposes a Trust model that combines peer profiling with anomaly detection technique. Each peer can establish trust based on its own prior activities with other peers by comparing the current activity of a peer with its historical data and Genetic Algorithm (GA).

Different from the traditional reputation models, this paper considers the reliability and trust value of nodes when evaluating the reputation value of nodes. On this basis, this paper presents a node reputation evaluation model based on time window, which will be the most recent node behavior. Give higher weights so that the behavioral trends of the nodes can be predicted more accurately. The PCA algorithm proposed in this paper is a defect of the previous method. It selects a trusted node for file copying based on the node reputation in the node with high query traffic, so that the file copy can actually serve the query initiating node, while reducing the copy overhead. Get high search efficiency.

3. Design

3.1. System model.
The relevant definitions of node reputation are given below:

**Definition 1** Business: The so-called transaction refers to the connection and communication behavior between nodes for the purpose of accessing resources such as data and services.

**Definition 2** Node Reliability: The reliability of a node to a node is derived from its transaction history.
\[ R_{u,v} = \frac{S_{u,v} - UnS_{u,v}}{N_{u,v}} \] (1)

Where \( N_{u,v} \) represents the number of transactions that occurred on node \( u \) and node \( v \). \( S_{u,v} \) is the number of transactions that are satisfactory, and \( UnS_{u,v} \) is the number of unsatisfactory transactions.

Node reliability is a measure of how much a node trusts another node, that is, how well a node considers a transaction with another node to achieve the desired goal.

**Definition 3** node credibility: The credibility \( T_{u,v} \) of node \( u \) versus node \( v \) is derived from the history of its recommendation information.

\[ T_{u,v} = \frac{ST_{u,v} - UnST_{u,v}}{NT_{u,v}} \] (2)

Where \( NT_{u,v} \) represents the number of recommended information provided by node \( v \) for node \( u \). \( ST_{u,v} \) is the number of recommended transactions, and \( UnST_{u,v} \) is the number of recommendations that are not satisfactory.

Node credibility is a measure of the trustworthiness of one node to another, indicating how much the former relies on the latter for providing recommended information.

The reputation of a node is a measure of the possibility that a node provides a high quality of service. For the evaluation of node reputation, this paper comprehensively measures the reliability and credibility of nodes. Since the node can improve its reliability instantaneously through a fake service in a short period of time, the reliability of the node is a local trust value, and it is also necessary to refer to the trust value of the node in the global node, that is, refer to the recommendation information of other nodes. However, while referring to the information given by other nodes, it is also necessary to consider the recommendation credibility of the recommendation node itself to decide whether to trust its recommendation. In fact, the reliability of nodes with high reliability is not necessarily high. For example, a node with high credibility may submit false recommendation information to its competing node; a node with high credibility may not necessarily have high transaction reliability if the node cannot provide reliable service.

**Definition 4** node reputation: the reputation \( P_{u,v} \) of the node \( u \) to the node \( v \) is derived from the weighted sum of the node reliability \( R_{u,v} \) and the credibility \( T_{u,v} \).

\[ P_{u,v} = \alpha \times R_{u,v} + \beta \times T_{u,v} \] (3)

\( \alpha > \beta \), and \( \alpha + \beta = 1 \).

In the network, we maintain an attribute group (ID, Reliability, and Credibility) for each node \( u \), where ID: network node ID.

3.2. Strategy

The node becomes the node to be copied, and the following two conditions must be met: 1) The query rate on the hotspot file \( f \) on the node is very large. Since our replication strategy is to copy the file to a node with a high query rate, the query message can be encountered before the target node is reached to shorten the search length. Therefore, the nodes with high query rate are first selected. We call these nodes a key node. 2) The reputation of the node is very high. Since the key nodes do not necessarily provide good services for other nodes, we select the nodes with high reputation as the key trusted nodes, so that the copy of the files deployed on them must be available, reducing the copy overhead and improving Service satisfaction rate and copy utilization.

**Definition 5** Key node: Assume that the query rate for file \( f \) on node \( i \) is \( q_i' \), the average query rate of file \( f \) is \( q_f ' \), if \( q_i' > q_f' \), then node \( i \) is the key node.
Where the average query rate of file $f$ is $q_f = \frac{\sum q'_f}{k}$, $k$ is the number of nodes participating in the query in the network.

Node $i$ periodically determines whether it is a critical node. Once it finds that its query rate exceeds the average file query rate, it sends a file copy request $r$ when forwarding the query message to the next hop node.

When the target node $j$ is overloaded, when the file copy request sequence $R = \{r_1, r_2, L, r_k\}$ initiated by the $k$ key nodes is received, the reputation survey request for the node $l$ is immediately initiated in the network. When a node in the network receives this request, it checks if it has a direct transaction with $l$. If there is no transaction or the transaction credibility with node $l$ is less than a predetermined threshold $\tau$, then the request is ignored, otherwise, this information is sent to node $j$. Node $j$ collects this information, combines its own information, calculates the reputation value of node $l$, and finds the key trusted nodes from these key nodes.

Node $i$ calculates the reputation of these $k$ nodes according to definition 4, and derives the reputation sequence of the node as $P = \{P_i, P_2, L, P_k\}$. We will find the key trusted nodes from $P$.

**Definition 6** Key trusted node: Assume that node $i$ is a key node, node $i$ has a reputation for node $j$ as $P_{ij}$, and if $P_{ij} > \delta$, node $i$ is a key trusted node. Where $\delta$ is the reputation threshold set by the system.

According to definition 6, we can find out the key trusted nodes in it, get $p'$, and the corresponding set of key trusted nodes is $c$.

Therefore, when a heavy load node receives a copy request initiated by multiple key nodes, it first filters out the key trusted node sequence, and then selects the appropriate node for file copying. So, according to what kind of rules, choose a node from $C$, here we use the cost model to calculate the benefits of copying files to nodes. We hope to get a large profit with a small number of copies.

Since it is desirable to minimize the number of copies, it is necessary to consider how many copies should be set to carry the partial query rate of the overloaded node to solve the hotspot problem. Assuming that the current query rate of the overloaded node $j$ is $q'_j$, and the maximum query rate that can be processed is $q'_{\text{max}}$, the node only needs to let other nodes share the file query rate $L = q'_j - q'_{\text{max}}$ to alleviate the hotspot problem.

Assuming that the distance from node $l$ to file owner $j$ is $d_l$ hop, the consumption of one hop is uniformly set to $b$, and the reputation of node $l$ providing service is $p_{lj}$, then the network consumption that may be saved by copying file $f$ to node $l$ is $q_f \times d_l \times b \times p_{lj}$. Assuming that the size of the file $f$ is $c_f$, and the copy overhead per unit file is $\gamma$, the cost of copying the file to the node $l$ is $\gamma c_f$, and finally the gain obtained by setting the file copy $f$ on the node $l$ is

$$e_l = q_f \times d_l \times b \times p_{lj} - \gamma c_f$$

Then the question is translated into:

Knowing the key trust node set $C$ of the overload node $j$ and the node reputation value $P' = \{p_{j1}, p_{j2}, L, p_{jm}\}$ corresponding to the node set, if the benefit brought by each node in $C$ is $E = \{e_1, e_2, L, e_m\}$, then a minimum node subset needs to be found from $C$, so that The benefit is just able to alleviate the overload $L$ of node $j$, that is,

$$e_1 x_1 + e_2 x_2 + L + e_m x_m \approx L$$  \hspace{1cm} (5)
Among $x_i = \begin{cases} 1, \text{replicate to } i \\ 0, \text{no replication} \end{cases}$, and $\sum_{i=1}^{n} x_i$ is the smallest.

### 3.3. Algorithm
Below we give the PCA copy replication algorithm.

Input: file copy request node set $C$ and node reputation set $P = \{p_{j,1}, p_{j,2}, \ldots, p_{j,n}\}$, copy target file $f$, overloaded node overload $L$, copy target node set $D = \emptyset$.

1. For the nodes in $C$, calculate the revenue of each node according to $P$ and formula (1), queue up in descending order of income (4), and obtain the revenue queue $E = \{e_1, e_2, \ldots, e_m\}$ corresponding to the node queue $S$ and the node.
2. Take the node $n_1$ of the queue head and add it to $D$, $D = D \cup n_1$, $S = S - n_1$.
3. Take the head element $e_1$ in $E$ and update the value of (5) $L$, $L = L - e_1$.
4. Judge $L < 0$? Set up turn 2, otherwise end.

Output: $D$ is the set of target copy nodes sought

### 4. Simulation
This experiment first constructs an unstructured P2P network. Generate a network topology that conforms to a power-rate distribution with a scale of 1000. Set the average degree of nodes to 5, and the search algorithm uses a random walk algorithm. Set the maximum forwarding distance of the query message to 7, that is, $\text{TTL} = 7$, the number of files is 200, the search interval is 1 period, and the maximum number of queries for a single node is 1. The capacity of the node conforms to the restricted Pareto distribution, which can truly reflect the network environment. Both the file requester and the requested file are randomly selected. The file query conforms to the Poisson distribution. We set the number of experimental cycles to 30. Finally, the EAD algorithm and the PAC algorithm are run to compare the copy hit rate with the search success rate. Each experiment is run 10 times and then the average value is taken as the experimental result.

![Figure 1](image-url)  
*Figure 1. Comparison of the utilization ratio of 5.*
As can be seen from Fig. 1 and Fig. 2, the replica hit rate, EAD is increased by about 10% compared with PAC, and the search success rate is increased by about 6%.

5. Conclusion

File replication technology for P2P networks is a very important research direction. Compared with the traditional file technology, the method can improve the utilization rate of the replica and the search efficiency by considering the influence of the free-riding node on the replica utilization.

Acknowledgement

This paper is supported by grant from the Science Research Project of Wuhan Business University (Grant No. 2015KA011)-Research on the Current Situation and Development Strategy of Chinese 3D Animation Films.

References

[1] C. L. Abad, "Load Balancing through Automated Replication in Unstructured P2P File Sharing Systems," 2007.
[2] D. Huang and Z. Hu, "Research of replication mechanism in P2P network," WSEAS Transactions on Computers, vol. 8, pp. 1845-1854, 2009.
[3] S. Mohammadi, et al., "An Enhanced Data Replication Method in P2P Systems," 2010.
[4] H. Shen, "An Efficient and Adaptive Decentralized File Replication Algorithm in P2P File Sharing Systems," Parallel and Distributed Systems, IEEE Transactions on, vol. 21, pp. 827-840, 2010.
[5] Yuh-Jzer Joung, Terry Hui-Ye Chiu, Shy-Min Chen, "Cooperating with free riders in unstructured P2P networks", Computer Networks 56 (2012) 198-212, 2012.
[6] Chithra Selvaraj, "Peer profile based trust model for P2P systems using genetic algorithm", Peer-to-Peer Netw. Appl. (2012) 5: 92–103, 2012.