A Distributed P2P Server System for Paper Sharing

Pingjian Zhang\(^1\), and Juanjuan Zhao\(^1,2\)

\(^1\) School of Software Engineering, South China University of Technology, Guangzhou, China
Email: pjzhang@scut.edu.cn
\(^2\) Nanchang Army College, Nanchang, China
Email: lzfjj@163.com

Abstract—P2P file sharing system is one of the hot research topics. However, most of such systems do not support auto extraction of metadata and provide only searches via resource titles. Combining the Chord algorithm and the SHA algorithm, this paper proposed a new P2P search model based on Distributed Hash Table (DHT). The procedures of establishing such P2P networks are presented. The design and implementation of the P2P Paper Sharing System is discussed. Experimental results show that the system achieves design goals.

Index Terms—P2P Search Model, P2P Server System, Metadata Extraction, P2P Paper Sharing System

I. INTRODUCTION

With the rapid development of information technology and the increase of the total volume of information, knowledge has played an essential role in the modern society. Knowledge management and sharing has become a hot research topic. As one of the most valuable knowledge resources, sharing of academic papers has important practical value, say, among research groups or developing teams. Being a resource sharing technology that has found a wide range of applications, the P2P networking \([1-3]\) provides a convenient means of realization for sharing academic papers freely, which is a nice supplement to those commercial digital libraries. So far, many P2P resource sharing projects have been carried out, for example, Napster \([4]\), Gnutella \([5]\), KaZaa \([6]\), Pastry \([7]\), Maze \([8]\) and Granary \([9]\), to name just a few.

Although the P2P technology has obtained high-speed development in recent years, there are still some key issues to be solved \([10-15]\). For example, the bandwidth occupation rate is high, the network expansibility is poor, and the resources usage is low, etc. The main reason lies in that resources in the P2P networks are of great dispersion, and nodes are free to join or exit, lacking unified and efficient management. How to effectively and reliably search resources in P2P networks becomes a challenging problem. In addition, most of the existing P2P resource sharing systems supports only searches by resource name, since resources are stored and manipulated in the whole document.

This paper presents a new P2P search model based on Distributed Hash Table (DHT), which consists of a layered server system. Ordinary server nodes (SN) form a chord ring, and a segment of SN of the ring is further managed by a super peep node (SPN), which is monitored by the coordinate node (CN). This layered model solves the problem of low search efficiency in the P2P network. To address the issue of metadata support when searching resources, the Chinese word segmentation module and the metadata auto extraction module are introduced. Experiments demonstrate that the new P2P paper sharing system works as desired.

The paper is organized as follows: in section 2, a new P2P search model is proposed along with the procedures to establish the P2P network. The design and implementation of the new P2P paper sharing system is discussed in section 3. Section 4 contains some final remark and conclusions.

II. A NEW P2P SEARCH MODEL

A. The P2P search model

The P2P search model consists of the following kinds of nodes:

- CN (coordinate node), that coordinates the resources managed by the SPN (super peep node) to prevent duplicate copies of resources.
- SPN, that is responsible for resources fetching, storage in FS (file server), and manage the join and exit of a group of SN (server node). SPN also reports routing information of SN to CN and broadcast routing information that it received from CN to its underlying SN.
- SN, that stores routing information about the whole P2P network and part of resource metadata in the (key, value) pairs. The SN form a chord ring.
- FS, that stores all the resources in the P2P network.

The topology for such a model is depicted in Fig. 1 below:

![Figure 1 The P2P search model](image-url)
B. The P2P networking strategy

The establishment of the P2P network consists of two steps:

1) Set-up of the SPN network

First of all, the CN starts up and initializes. Then, SPN registers to CN which generates a unique identifier and return it to the newly registered SPN. After receiving the registration response, SPN sends an acknowledgement to CN which will then notify all other SPN of the registration information. When the number of registered SPN arrive a certain threshold (which can be configured), the CN broadcasts a “finish” message to all SPN. The SPN then requests CN of routing information of all SPN and the CN responds with the required information. The process is described in Fig. 2 below.

2) Set-up of the SN network

Only after SPN network has been set up can the SN network begin construction. At this stage, SN registers to SPN, which generates and returns a unique identifier by hashing the ip address of the SN via the SHA-1 algorithm. After receiving the registration response, SN sends an acknowledgement to SPN which will broadcast the routing information of the new SN to other SPN and then notify all other SN under its management of the registration information of the new SN. The new SN, in turn, will request its SPN of the global routing table.

When the number of registered SN arrive a certain threshold (which can be configured), the SPN broadcasts a “finish” message to all SPN including itself. When the number of “finish” message a SPN received equals the number of all SPN. The SPN then broadcasts the “whole network finish” message to all SN under its management. The process is described in Fig. 3 below.

3) Storage of resources

When SPN receives resources uploaded by clients, it stores it in the correspondi ng FS, extracts metadata and hashes them using SHA-1. Each (key, value) pair is sent to the successive SN nearest to the key in the chord ring, together with the information about the FS.

4) Search of resources

Upon the request of a resource, the SN extracts and hashes the metadata to obtain the key. Then, it queries the successive SN nearest to the key in the chord ring. The destination SN returns information about the FS that stores the desired resource. The SN fetches the resource from the FS and returns to the client.

5) Join of a new SN

After the whole network is constructed, if a new SN needs to join the network, it proceeds quite the same way as in the steps of set-up of the SN network.

6) Exit of an SN

When an SN wants to exit, it sends an “exit” message to its SPN which forwards the message to other SPN and SN under its management. Then, the successive SN to the exiting SN pulls data from the exiting SN and stores them. Upon receiving the “transfer finish” message, the exiting SN deletes data and exits, as shown in Fig. 4.

Figure 2 Sequence diagram of set up of SPN network

Figure 3 Sequence diagram of set up of SN network

Figure 4 Sequence diagram of exit of an SN
III. DESIGN AND IMPLEMENTATION OF THE P2P PAPER SHARING SYSTEM

The system is designed to be a paper sharing system among research and development groups. Group members can upload valuable papers to share with others, and can search and download papers wanted.

The system consists of two subsystems: the P2P server system that falls into the P2P search model mentioned above, and the client end.

A. System analysis

The system is divided into 4 modules: the CN module, the SPN module, the SN module, and the client module, whose functionalities are list as follows.

The CN module will
1) handle registration of SPN and allocate a unique identifier for each SPN.
2) broadcast registration of new SPN to other SPN.
3) determine if the construction of SPN network is done.
4) assign SPN to clients.

The SPN module will
1) handle registration of SN.
2) broadcast routing information to other SPN and SN under its management.
3) store the global routing table.
4) process Chinese word segmentation, inverse document indexing and metadata hashing.
5) response to client query.
6) handle the exit of SN.

The SN module will
1) handle registration to SPN.
2) handle exit of SN.
3) store paper and metadata information.
4) synchronize with other SN.

The Client module will
1) extract metadata of papers and upload to SN together with the paper.
2) act as interface for searching papers via paper title and other metadata information.

B. Architecture of the system

Following the system requirement analysis, the system architecture consists of 4 subsystems and is interacts as in Fig. 5 below, while the deployment diagram shown in Fig.6.

C. Design and Implementation of the system

This subsection will highlight design and implementation issues of some key components.
1) The data dictionary class

Data dictionary is fundamental for the task of Chinese word segmentation. The class diagram is given in Fig. 7.

2) The communication module

This is one of system cores, responsible for the reliable transfer of data among various nodes and triggers the data processing in the business layer. The module contains two abstract classes and a helper class.

3) Chinese word segmentation module

Chinese word segmentation is a key task in the system that makes metadata extraction possible. Current system adopts the longest string matching approach. The module consists of two classes: MyBaseAnalyzer and SentencesAnalyzer.

4) The CN module

The main class in the module is the CenterMian class whose main() method will load system parameters via
class ConfigUtil and will create an instance of CenterServer, then start it.

5) The Client module

This module consists of functionality class such as BrowserClient, PdfboxParse, CenterRouter, Command, DestInfor, NormalMeta and SearchResult.

D. Experiments

A prototype Paper Sharing System has been developed and deployed for system test. The aim is to verify

1) The construction of SPN network. Configure the number of SPN to be 4 and register them to check if the routing table in CN is correct.

2) The maintenance of SN network. Start SPN, join some SN to each SPN, then, detach some SN, and join the exited SN to SPN again, to check if the routing table in the SPN is correct.

3) Search of papers. Search papers via a browser to check if papers can be located by title or metadata information correctly.

4) Prevention of single point of failure. When an SN failed, its resources are transmitted to its successor and are still available.

The system is deployed in a local network with 2 SPN and 4 SN which is depicted in Fig.13. The test cases listed above are run and results are recorded as below.

1) The construction of SPN network.

Start the service in CN with ip 125.216.250.70, then, start the services in two SPN with ip 125.216.249.109 and 125.216.249.200 respectively. The CN routing table reads as

<table>
<thead>
<tr>
<th>Ip</th>
<th>superId</th>
</tr>
</thead>
<tbody>
<tr>
<td>125.216.249.109</td>
<td>1</td>
</tr>
<tr>
<td>125.216.249.200</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 8 The communication module class diagram

Figure 9 The MyBaseAnalyzer class diagram

Figure 10 The SentencesAnalyzer class diagram
2) The maintenance of SN network.

Add an SN to SPN with superId 1, and 3 SN to SPN with superId 2. The SPN routing table reads as

<table>
<thead>
<tr>
<th>Ip</th>
<th>status</th>
<th>superId</th>
<th>hashkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>125.216.250.70</td>
<td>0</td>
<td>1</td>
<td>579269F26FA29E8B57E353D791A382E7E749A529</td>
</tr>
<tr>
<td>125.216.250.109</td>
<td>0</td>
<td>2</td>
<td>48F2D151C5D6A84E7E11A74E8FFBE17CDB543</td>
</tr>
<tr>
<td>125.216.250.125</td>
<td>0</td>
<td>2</td>
<td>11F7A4943F9956C7A3219339774C8098428</td>
</tr>
<tr>
<td>125.216.250.200</td>
<td>0</td>
<td>2</td>
<td>6EA3FAF06F0F1525F94214B54E5EB7D3374BB11</td>
</tr>
</tbody>
</table>

Shutdown the SN with ip 125.216.250.70, the corresponding status in Table 2 turns to be 1; restart it again, the status changes back to 0.

3) Search of papers. Search by title is carried out as blow.

<table>
<thead>
<tr>
<th>Ip</th>
<th>status</th>
<th>superId</th>
<th>hashkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>125.216.250.70</td>
<td>0</td>
<td>1</td>
<td>579269F26FA29E8B57E353D791A382E7E749A529</td>
</tr>
<tr>
<td>125.216.250.109</td>
<td>0</td>
<td>2</td>
<td>48F2D151C5D6A84E7E11A74E8FFBE17CDB543</td>
</tr>
<tr>
<td>125.216.250.125</td>
<td>0</td>
<td>2</td>
<td>11F7A4943F9956C7A3219339774C8098428</td>
</tr>
<tr>
<td>125.216.250.200</td>
<td>0</td>
<td>2</td>
<td>6EA3FAF06F0F1525F94214B54E5EB7D3374BB11</td>
</tr>
</tbody>
</table>

with superId 2, the SPN routing table reads as

And search again, the paper still exists in the network, although in another SN (refer to Fig. 16).

IV. CONCLUSIONS

P2P network is a promising technology for resource sharing system for group/team/organizations. Most of current systems suffers high bandwidth pressure and/or low search efficiency. In this paper, a layered P2P server system model based on the Chord and SHA algorithm is proposed to resolve the high bandwidth problem. The introduction of CN eliminates unnecessary duplicate resources storage in the network; the introduction of SPN eliminates the flooding of routing information around the network. By incorporating Chinese word segmentation and metadata extraction modules, the system supports versatile and convenient search approaches, thus enhancing search efficiency. Furthermore, the design and implementation of a P2P Paper Sharing System based on the new model is discussed. Experimental results demonstrate that the system achieves its design goals.

There remain some topics uncovered by the present paper. Semantic search, for example, is an important issue that is not discussed here and deserves further study. These will be addressed in subsequent papers.

REFERENCES


