Research of Media Material Retrieval Scheme Based on XPath

Shuang Feng[1], Weina Zhang[2]

[1]School of Computer Science,Communication University of China, Beijing, China
fengshuang@cuc.edu.cn
[2]Computer NIC Center,Communication University of China, Beijing, China
zhangweina@cuc.edu.cn

Abstract. With more and more media materials appear on the internet, it comes a sharp problem of how to manage these resources and how to search them efficiently. We construct a management system of media material for the reliable wideband network. According to a detailed analysis of the characteristic of media material queries, we proposed a hierarchical indexing mechanism based on XPath language to discover the resources that match a given query. Our system permits users to locate data iteratively even using scarce information. The description of materials is mapped onto the DHT index. Our indexing scheme has good properties such as space efficient, good scalability and resilient to arbitrary linking.

Keywords: P2P; XML;XPath; media material retrieval

1 Introduction

With the public's enthusiasm for video creation continues to increase, they tend to create and share their works through internet. Personal media production and distribution not only need production tools, but also need mass media materials. The establishment of shared material library through internet is an effective way to solve this problem. Taking into account of the enormous amount of media materials, we can use a distributed network storage system for storage and services. We also need an efficient indexing technology to allow users retrieve the materials accurately and quickly.

2 Related Work

A P2P (Peer to Peer) network is a distributed system. According to the topology of the network, the P2P system can be divided into two kinds: unstructured P2P networks and structured P2P networks.

Inefficient routing and bad scalability are the main shortcomings of unstructured P2P networks. In contrast, the structured P2P network with DHT (Distributed Hash Table) has good scalability, robustness, and can provide accurate location, the rate of search success is also high. However, the DHT technique can only support accurate retrieval

and can't support the complex retrieval such as semantic retrieval. The users usually do not know the exact materials they want, they often tends to retrieve the materials by category and interested in the classical materials with a higher utilization rate in that field. So DHT index can't be used directly in media semantic retrieval .

Currently, the research of semantic retrieval based on DHT includes: [1] gets all the results of each keyword and then find the intersection, but because this method requires to transfer a large number of middle documents, it will consume a lot of network band. [2] supports multi-keyword search by using vector space model. [3] uses latent semantic indexing to overcome the affect of synonyms and noise in VSM . [4] builds a theme overlay network to improve the rate of full search. [5] designs a multi-attribute addressable network to support multi-attribute queries and range queries. Most of the above methods analyze the communication cost and retrieval accuracy, but for the particular application fields such as material retrieval, in most cases, users may only provide partial information, therefore, it is essential to build a relation mapping content-based index onto DHT index.

3 Structure of Large-scale Material Retrieval System

One of the most important researches in P2P retrieval is structured P2P network retrieval based on DHT. Some of the DHT retrieval algorithms such as Chord[6],CAN[7] are designed for fluctuations in the worst-case network. The problem of this kind of work is long delay time. For the reliable wideband network with server clusters, we use a method called distributed storage and centralized retrieval to improve the efficiency of retrieval. Accordingly, the retrieval and management system of large-scale media material is shown in Figure1. It contains the following parts.

- Catalog server: responsible for labeling the materials according to the international standard of media description.
- Mapping server: responsible for mapping media description onto DHT index.
- Retrieval server: help users find relevant materials through a variety of ways. Sub-retrieval is used to share concurrent search pressure.
- Directory server: responsible for managing user's virtual directory and searching the exact location of block-stored materials. It also provides a directory query in DHT way.
- Storage node: the storage space is composed of many distributed storage nodes. They provide services for materials with high demands such as block storage and other services.
- Client: The client may also constitute a storage space for those materials with relative lower demands. And as an agent, the client is used to access the whole storage space and submit media materials to catalog server.

Before submitting a material, the user needs to fill in relevant information about the material and point out both the share scope and the way of storage (local storage or distributed storage).Then the catalog server will label the material according to international standards. Thus, a standard catalog file is produced and the catalog file is submitted to mapping server. The mapping server will map the description of mate-

rials onto DHT index and save this relationship to the retrieval server. The retrieval server accepts the user's search requests and returns relevant query results. Then the user chooses one of the results and the client will tell the storage system which material the user chooses. The storage system searches its own directory server and returns a set of IP address that stores that material. Finally, the client downloads the material by P2P.

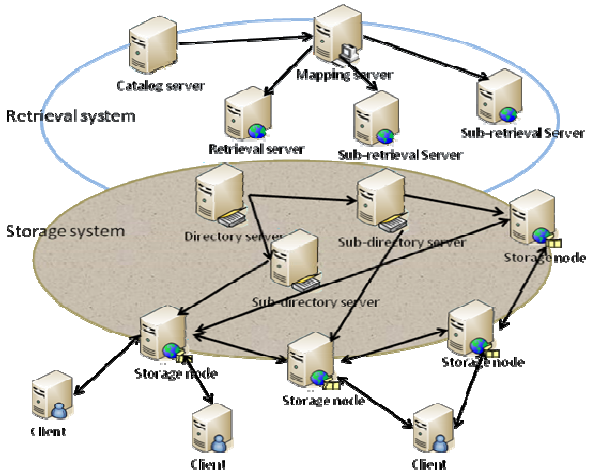

Fig. 1. Structure of large-scale materials retrieval system

4 Services of Large-scale Material Index

How to build an efficient index to help users find what they want quickly among large-scale materials? Relative works about the management of large-scale distributed materials includes P2P-based file system(OceanStore [8],Tsinghua Granary [9] systems) and distributed indexing(Chord, CAN and Pastry [10]).However, the main problem is that if we use DHT directly on the retrieval of semantic materials, we can't support complex queries such as semantic queries. But in many cases, users may only provide part of information that they want search. Therefore, it is essential to map contend-based index onto DHT index.

4.1 Mapping Description of Materials onto DHT Index

To solve the problem of large-scale material retrieval, our system is organized hierarchically. As shown in Figure 2. Here are the roles of these three layers. Media description mapping layer maps the description of media onto the DHT-based index. The index of distributed media materials are organized by distributed index layer. We use Pastry, a kind of widely-used distributed hash table technology, to compute in-

dexes and route objects. Storage layer is used to store media materials. The management of distributed content-based media resources is done by this architecture.

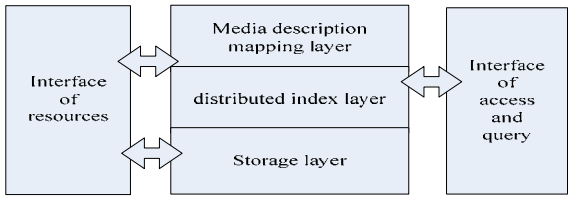

Fig. 2. System architecture of large-scale materials retrieval

In the P2P file sharing systems, file search is based on file names and file names can be seen as a brief description of the file. Further expanding this point of view, we can create media material description with XML. Data query is based on these descriptions. Figure 3 shows the XML description of the media file.

```
<program>
  <director>
    <LastName>Zhang</LastName>
    <FirstName>Yimou</FirstName>
  </director>
  <title> Red Sorghum</title>
  <type>movie</type>
  <date>1987</date>
  <format>general</format>
</program>
```

a) document d1

```
<program>
  <director>
    <LastName>Zhang</LastName>
    <FirstName>Yimou</FirstName>
  </director>
  <title> 2008 Olympic opening ceremony</title>
  <type>art</type>
  <date>2008</date>
  <format>wide</format>
</program>
```

b) document d2

```
<program>
  <director>
    <LastName>Lang</LastName>
    <FirstName>kun</FirstName>
  </director>
  <title> 2009 Spring Festival Party</title>
  <type>art</type>
  <date>2009</date>
  <format>general</format>
</program>
```

c) document d3

Fig. 3. Descriptions of document d1,d2,d3

To query a material based on XML description, we can use XPath language. XPath uses path expressions to select nodes or node sets in XML document. A complete query expression is an expression with all the nodes in XML documents which are not empty. as shown in Figure 4. $q1$ is the complete query expression of document d1 in Figure 3. When the complete query expression is processed with hash functions, we can create DHT index. In this way, not only the original semantic information can be preserved, but also the materials can be stored correctly to the distributed storage system. Thus the description of media materials can be mapped onto DHT index.

```
q1= / program[ director [ Last Name/Zhang ]][Title/Red Sorghum]
     [ Type/Movie]  [Date/1987] [Format/general]
q2= / program[ director [ Last Name/Zhang ]] [ Type/art]
q3= / program/director [ Last Name/Zhang ][First Name/Yimou]
q4= / program/Title/Red Sorghum
q5= / program/ Type/art
q6= / program/ Last Name/Zhang
```

Fig. 4. Examples of queries

4.2 Description of Hierarchical Index

In the management system of large-scale media material, the materials are stored in a distributed network storage, but in order to achieve fast retrieval response time, we adopt a centralized retrieval method by using a cluster of retrieval servers. Now we do some definitions:

Given two queries q and q', if all the results returned by query q are included in the results returned by query q', then we said query q' contains query q, denoted by $q' \supseteq q$. Particularly, if query q is the complete query expression of document d, denoted by $q \equiv d$. Figure 5 is a partially ordered tree of Figure 4.

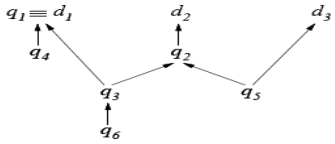

Fig. 5. Partially ordered tree of Figure 4

According to complete query expression, we can build a query expression set for all the queries which are frequently searched. $Q = \{q1, q2, ...qn\}$,where $qi \supseteq q$, we stored $< qi, q >$. For example, according to the document set shown in Figure 3, we can establish a hierarchical index shown in Figure 6. Each rectangle represents an index entry and each index entry stores the corresponding information. Top-level index entry represents a complete material description and other index entries maintain a relation between query conditions. In this way, we can provide an iterative way to help uses find what they want.

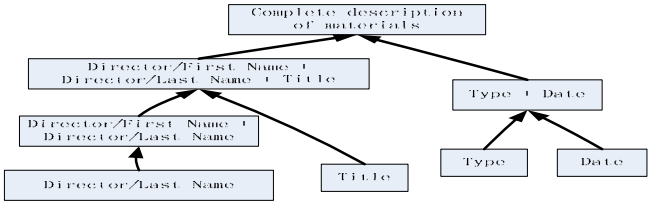

Fig. 6. Example of indexes

4.3 Procedure of Retrieval

If query $q0$ is a complete query expression, we will get the description of the corresponding document directly, or we will get a query sequence $\{q1, q2, ...qn\}$, where $q0 \supseteq qi$ Then the user can choose a query that he is interested as a new query, continue to iterate until he finds the necessary resources. For example, for the query $q6$ given in Figure 4, query $q3$ is found first. Then we get document d1 and d3 through

$q3$. When the given query $q0$ is not included in index entries, but there were some documents that satisfied the query condition. Then we find $qi \supseteq q0$, where qi is the current level of an index entry..

5 Conclusion

The distributed wideband collaborative service environment based on WAN provides technical support for carrying out more and better multimedia production business. This paper builds a retrieval and manage system of large-scale media materials. According to the unique characteristics of materials, we mapped the description of media onto the DHT index and proposed hierarchical indexing mechanism to improve query efficiency. Hierarchical indexing mechanism describes the indexes based on XPath, it helps users find the materials in an iterative way.

Aknowledgements

The paper is supported by the National Key Technology R&D Program of China (2012BAH02F04), and CUC Engineering Planning Project (XNG1125).

References

1. Zhou Feng, Zhuang Li, Zhao B Y, et al. Approximate Object Location and Spam Filtering on Peer-to-peer Systems. In: Proceeding of ACM/IFIP/USENIX Intl Middleware Conference, Middle-ware,2003
2. Tang Chunqiang,Xu Zhichen,Dwarkadas S.On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems[C] SIGIR'04.
3. Tang CQ, Yu ZH, Mahalingam M. pSearch: Information retrieval in structured overlays. ACM SIGCOMM Computer Communication Review, 2003
4. Xianghua Fu,Research on Construction and Searching Algorithms of Topic Overlay Networks,Computer Science, 2007,6
5. M Cai,M Frank,J Chen,et al.MAAN:A multi-attribute addressable network for grid information services[J].Journal of Grid Computing, 2004
6. I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In Proc. ACM SIGCOMM ,2001.
7. S. Ratnasamy, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In Proc. ACM SIGCOMM, 2001.
8. Kubiatowicz J, Wells C, Zhao B, Bindel D, Chen Y, Czerwinski S, Eaton P, Geels D, Gummadi R, Rhea S. OceanStore: An architecture for global-scale persistent storage. In: Proc. of the 9th Int'l Conf. on Architectural Support for Programming Languages and Operating Systems. 2000. 190-201.
9. Zheng, W, J Hu, and M Li, Granary: architecture of object oriented Internet storage service. E-Commerce Technology for Dynamic E-Business, 2004. IEEE International Conference on, 2004: p. 294-297.
10. A. Rowstron and P. Druschel. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In Proceedings of Middleware, Nov 2001.