Chapter 8
An Approach for Large Scale Retrieval Using Peer-to-Peer Network Based on Interest Community

Shuang Feng, Shouxun Liu and Yongbin Wang

Abstract Conventional multimedia information retrieval systems use a central system to store and index multimedia data. Inherent limitations of such a central approach surface many problems, such as insufficient bandwidth, server overloading and failures. In order to retain the original system and control the cost, the paper shares the access pressure of central servers by constructing a peer-to-peer network based on interests. A user's interest is computed by mining this user's search behaviour periodically. Then researcher form a peer-to-peer network based on interests by clustering peers with similar interests. Centralized server will push relevant multimedia information to certain communities on time. By this way, uses can further clear what they want, and useless retrievals on servers will be dramatically decreased. The experimental results evaluate the average search path length of unstructured P2P network, semi-distribution P2P network and the P2P network based on interest community and demonstrate the efficiency of the approach.

Keywords Interest community · Peer to peer · Multimedia retrieval · User profile

S. Feng (✉) · Y. Wang
School of Computer Science, Communication University of China, Beijng, China
e-mail: fengshuang@cuc.edu.cn

Y. Wang
e-mail: ybwang@cuc.edu.cn

S. Liu
The Graduate School of Communication University of China, Beijng, China
e-mail: sxliu@cuc.edu.cn

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_8,

8.1 Introduction

With the rapid development of information technology, multimedia applications have been widely used in people's daily lives. The need of developing effective and efficient multimedia information retrieval technologies has been identified in recent years. Due to the large amount of computational power needed for the searching and processing of multimedia data, distributed multimedia information retrieval has attracted researchers' attentions [1]. But there are still many original retrieval systems with central servers. As more and more multimedia objects are collected and the scale of the applications grows, the inherent limitations of such a central approach surface, such as insufficient bandwidth, server overloading and failures [2].

The rapid development of P2P technology made it as one of the most disruptive tools for the construction of large-scale distributed system over Internet. P2P adopts a distributed and decentralized architecture, and each peer of P2P is equal and acts as both a server and a client, so P2P removes the drawback of the structure of central server.

Compared with breadth-first search in the network, retrieval efficiency can be greatly improved if peers which most probably contain relevant data are visited first. The assumption is that every peer has its own topics of interest. These interested topics are the reflection of the interests of the user behind the peer. The user is also more likely to query multimedia data on the topics that he/she is interested in. The multimedia information retrieval process can be used to facilitate relation establishment between peers in the network. The relevancy judgment of the results returned made by the querying peer can be seen as an assessment to the information retrieval performance of their corresponding data sources. In this way, we model the multimedia information retrieval network as a social network and propose a multimedia retrieval model based on P2P and a recommendation system based on interests to decrease the access of central servers.

The remainder of this paper consists of five parts. In Sect. 8.2, related work is introduced. In Sect. 8.3, an interest-based P2P system architecture is presented. Section 8.4 describes the calculation of user profile. Section 8.5 shows algorithm of community evolution. Section 8.6 is the evaluation of our approach. Finally, conclusions are presented in Sect. 8.7.

8.2 Related Work

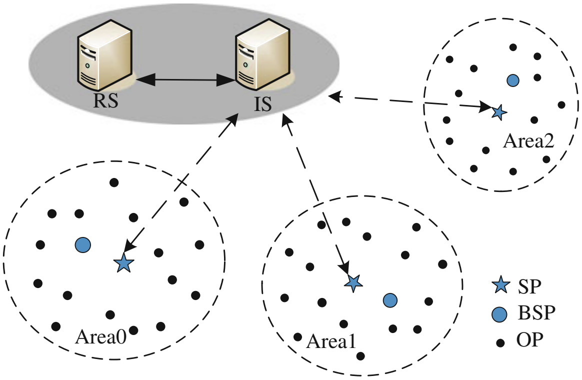
To reduce the number of query search messages in the P2P network, many algorithms propose the concept of groups or clusters in P2P networks [3–5]. The nodes with more powerful resources (in term of processing power, memory and bandwidth) are the suitable candidates for the role of server, whereas, less powerful nodes become clients [4]. A new protocol was proposed for building and repairing of overlay topologies based on the formation of interest-based superpeers [5]. An interest-based superpeer algorithm creates groups or societies that have common interests. A semantic community establishing method based on consideration of semantic similarity degree is proposed, trust degree and active strength, which improves the structural and resource-located veracity of P2P resource organization [6]. SOSPNET maintains a superpeer network topology that reflects the semantic similarity of peers sharing content interests [7]. Superpeers maintain semantic caches of pointers to files, which are requested by peers with similar interests. Client peers, on the other hand, dynamically select superpeers offering the best search performance. There has been work in community construction in P2P networks [8, 9]. Four sources of self-construction of P2P communities, namely, ontology matching, attribute similarity, trust, and link analysis were introduced to form community [8]. In communities for sharing academic papers, each peer computes its trust in the peers with whom it interacts [9].

8.3 Superpeer Overlay Network

To take full advantage of the client's ability, we build a P2P overlay network on internet to alleviate the pressure on servers. Relevant peers are clustered based on their historical behavior on the retrieval server, called interest community. By this way, the network is divided into different interest communities. Each of community elects a superpeer and a backup superpeer, they are responsible for cooperative management of their own community. IS will sent heartbeat information on time so that it knows weather communities work well. SP will also sent heatbeat information to IS if there have some changes in the community. IS will push recommendation multimedia information to certain communities on time. By this way, uses can further clear what they want and useless retrievals on servers will be dramatically decreased. The Architecture is shown in Fig. 8.1:

There are five entities in our system, they are:

Fig. 8.1 Overall system architecture of the proposed scheme



- RS (Retrieval Server): Store all the information of multimedia materials, responsible for generating commendation file and responding client requests for the retrieval.
- IS (Index Server): Create and maintain community information dynamically.
- SP (Superpeer): Each community in P2P overlay network elected a superpeer based on the capacity of the node dynamically. SP is responsible for getting commendation file from IS and maintaining its community work well.
- BSP (Backup Superpeer): In order to avoid the problem of single point failure, SP designated a backup superpeer based on the capability and reputation of nodes in its community, BSP maintained the community information with SP cooperatively. When SP can't serve the community, BSP will take over the community and start updating mechanism.
- OP (Ordinary Peer): OP can join or exit the retrieval network. Each OP stores the information of RS, IS and community details such as SP and BSP.

8.4 User Profile Construction

Personalized social search can be achieved by utilizing historical query data from people in a community with similar interests.

Definition 1 Assuming each multimedia file has a topic and the number of topics is limited, the similarity of peers is defined as the similarity of a set of weighted topics.

$$\begin{aligned} Sim(P_1, P_2) &= Sim(<T_i, \lambda_i>, <T_j, \lambda_j>), i = 1, 2, \ldots\ldots m, j = 1, 2, \ldots\ldots n \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} Sim(T_i, T_j) \times (\lambda_i, \lambda_j) \end{aligned} \tag{8.1}$$

where P_i is the peer, T_i is the topic, λ_i is the weight of T_i, it can be calculated as follows:

$$\lambda_i = \frac{N_i}{\sum_{j=1}^{n} N_j} \tag{8.2}$$

where N_i is the number files belongs to T_i among all the files.

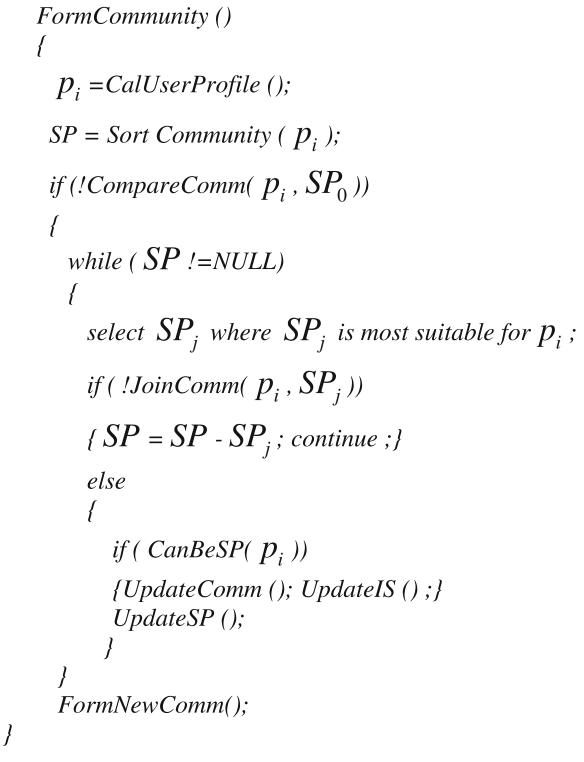
Many methods have been developed to measure the similarity of concepts. Our content summary of a peer is calculated based on the method proposed by Yuhua Li et al. [10], as follows:

$$Sim(T_i, T_j) = f_1(l) f_2(h) = e^{-\alpha} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{8.3}$$

where l is the shortest path length between topics, h is the depth of subsumer in the hierarchy semantic nets, α and β are parameters scaling the contribution of shortest path length and depth, respectively.

8.5 Algorithm of Interest-Based Community Evolution

The algorithm, which is used to build the relationship between interest-based superpeer and clients, is shown in the Fig. 8.2. When the user searches multimedia on central servers for the first time, user's interaction will be recorded. By analyzing these, we can calculate the user's profile P_i. According to P_i, IS will sort all the communities through formula (3), mentioned in user profile construction. Compared P_i with SP_0, if there is no change, it means community SP_0 is still the most suitable one for P_i. Otherwise, we select SP_j where SP_j is most suitable for P_i. P_i asks for joining community SP_j. If it is OK, the community SP_j will check if P_i is able to be new SP according to its capacity. If so, both the community and IS status will be updated. If P_i can't join SP_j due to connection problems, IS will choose another SP for P_i until SP is NULL. If SP is NULL, a new community will be created.



```
FormCommunity ()
{
   $p_i$ =CalUserProfile ();
  SP = Sort Community ( $p_i$ );
  if (!CompareComm( $p_i$ , $SP_0$ ))
  {
    while ( $SP$ !=NULL)
    {
      select $SP_j$ where $SP_j$ is most suitable for $p_i$ ;
      if ( !JoinComm( $p_i$ , $SP_j$ ))
      { $SP = SP - SP_j$ ; continue ;}
      else
      {
         if ( CanBeSP( $p_i$ ))
         {UpdateComm (); UpdateIS () ;}
         UpdateSP ();
      }
    }
   FormNewComm();
}
```

Fig. 8.2 Algorithm of interest-based community evolution

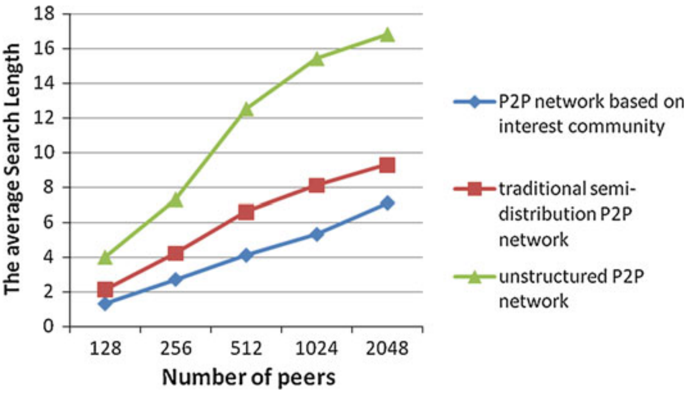


Fig. 8.3 The average search length of three kinds of P2P networks

When the system extends, the retrieval server must adjust two parameters, $K_{\max}$ (the max number of communities) and $N_{\max}$ (the max number of peers in a community). By merging or partitioning the communities, the communities can be adjusted dynamically, so that the scale of communities is average and each peer is connected to the community with common interests. The RS can control the network scale and the communities' partition, and enhance the controllability and availability of P2P overlay network.

8.6 Evaluations

We use P2PSIM as our evaluation tool. P2PSIM is a P2P simulation on Linux which can simulate kademlia, chord for P2P, and can simulate more P2P protocols by extending protocols and verify their functionalities. In our experiments, we generated 128, 256, 521, 1024, 2048 ordinary peers separately. Then we assigned ten queries to each peer randomly. Based on these queries, the similarity of each peer can be calculated according to formula (3). We also assume that peers with common interests share similar files. The max number of peers in each community is 50. Then we set a value to each peer that represents the capacity of the peer. Superpeer and Backup Superpeer are elected based on the capacity. All the results are the average value of five separated experiments. Figure 8.3 shows the average search path length of unstructured P2P network using flooding algorithm, traditional semi-distribution P2P network and P2P network based on interest community. From the figure, we can find out that the average search path length of our approach was dramatically decreased.

8.7 Conclusion

High-load and high-concurrency ask for a very high capacity of server. In this paper, authors proposed an approach for large scale retrieval by using peer-to-peer network based on interest community. Social information is acquired from a social network service application. Members in a community share certain interests. Communities are under the control of superpeers and backup superpeers. The central server will send relevant recommendation information to communities so that the users can further clear what they want and reduce useless retrieval dramatically.

Acknowledgments This paper is supported by the National Science and Technology Support Project (2012BAH02F02-01) and the National High Technology Research and Development Program of China (2011AA01A107).

References

1. Xia, T., Wang, F., Liu P., Palanivelu S.: Managing and searching distributed multi-dimensional annotations with large scale image data. In: Proceedings of the International Workshop on Multimedia Content Analysis and Mining (MCAM 2007), vol. 4577, pp. 361–370. LNCS, Springer (2007)
2. Rasolofo, Y., Abbaci, F., Savoy, J.: Approaches to collection selection and results merging for distributed information retrieval. In: Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, ACM, pp. 191–198. Atlanta, Georgia, 5-10 Nov (2001)
3. Gatani, L., Lo Re, G., Gaglio, S.: An adaptive routing protocol for ad hoc peer-to-peer networks. In: Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, pp. 44–50 (2005)
4. Montersor, A.: A robust protocol for building superpeer overlay topologies. University of Bologna, Bologna, Technical Report, UBLCS- 2004-8 (2004)
5. Ashraf Khan, S.K., Tokarchuk, L.N.: Interest-based self organization in group-structured P2P networks. In: Proceedings of the 6th IEEE Consumer Communications and Networking Conference, pp. 1–5 (2009)
6. Wang, Li, Hu, G.-X.: P2P semantic community model based on interest and trust evaluation. Comput. Eng. **35**(13), 11–13 (2009)
7. Garbacki, P., Epema, D.H.J., van Steen, M.: The design and evaluation of a self-organizing superpeer network. IEEE Trans. Comput. **59**(3), 317–331 (2010)
8. Liu, K., Bhaduri, K., Das, K., Nguyen, P., Kargupta, H.: Client-side web mining for community formation in peer-to-peer environments. SIGKDD Explorations **8**(2), 11–20 (2006)
9. Wang, Y.: Trust-based community formation in peer-to-peer file sharing networks. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), pp. 341–348. IEEE Computer Society, Beijing, 20–24 Sept 2004
10. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. **15**(4), 871–881 (2003)