A Document Recommendation System Based on Clustering P2P Networks

Feng Guo[1] and Shaozi Li[2,*]

[1] Dept. of Computer Science, Xiamen University, Fujian, China, 361005
betop@xmu.edu.cn
[2] Dept. of Computer Science, Xiamen University, Fujian, China, 361005
szlig@xmu.edu.cn

Abstract. This paper presents a document recommendation system based on clustering peer-to-peer networks. It's an unstructured P2P system. In this system each agent-peer can learn user's interest, then it helps user share and recommend documents with the other users. Since each peer in our P2P networks is a node, in order to cluster them, we import the concept of Group. Each group is composed of peers. The types of documents, which belong to a same group, are uniform. This paper presents how these peers help users to share and to recommend documents, and how they cluster into groups. Our experiment results show the advantages of the document recommendation system.

Keywords: Recommendation System, Clustering P2P, Reputation Management.

1 Introduction

Nowadays information in the internet is quickly increasing, which promotes the development of text mining and information retrieval techniques such as file sharing systems based on P2P networks. These systems have gained wide popularity. Some reports [1] suggest that P2P traffic is the dominant consumer of bandwidth ahead of Web traffic. But these systems still have some problems in querying. The users can only query files by file-names, which is too simple to express their contents. This paper presents a document recommendation system based on clustering P2P networks, called hybrid filtering system, which uses document's content and user's opinions to decide the relationship between document and user. It has higher precision in querying than other file sharing systems.

Content-based information filtering can recommend documents effectively by users' interest which has been known before. While collaborative filtering can learn users' new interest from other similar users, but it suffers from cold start [2] problem. The core of filtering algorithm in our system is a hybrid document filtering algorithm, including content-based filtering and collaborative filtering.

* Corresponding author.

Y. Luo (Ed.): CDVE 2007, LNCS 4674, pp. 226–233, 2007.

In this paper we import the concept of Group, which is composed of peers. The types of documents, which belong to a same group, are uniform. Peers can calculate the similarity with each others. Then the similar peers can cluster automatically into a same group. But the "similar user" is very difficult to decide in an unstructured P2P networks without a central server [3]. So we develop a new reputation measurement in our system, which is called authority score and used to represent group's property. Moreover our system applies recommendation threshold instead of N-best in hybrid filtering.

This paper is organized as following. Section 2 provides the structure and advantage of our system. Section 3 describes the recommendation algorithm and the reason we use threshold instead of N-best. Section 4 describes the usage of authority score and the calculation algorithm. Section 5 presents the clustering algorithm of peer nodes. Section 6 is our experiment and conclusion.

2 Structure of the P2P Networks

Our system is an unstructured P2P networks based on JXTA, which is an open source P2P platform. Every peer in the system can join or leave the discussion group freely. They can not only share and search documents but also obtain recommendations of documents. Every peer can act as a rendezvous peer when necessary, typically when it is a gateway. The discussion group is a virtual community which allows users to join and share documents. The group information is spread and held by peers who have joined the group and have been the rendezvous peers in the group. The structure of our system is shown in Figure 1.

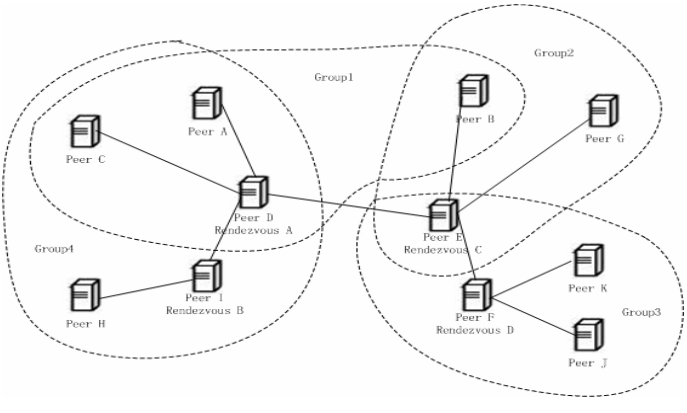

Fig. 1. Structure of the P2P networks

3 Recommendation Algorithm Based on Threshold

We use recommendation threshold in our system instead of the traditional N-best recommendation. The N-best recommendation algorithm works well in traditional systems, because in the C/S model the servers contain all the items and the number of items user may like is always far more big than "N". But in the unstructured P2P