TRES-CORE: Content-Based Retrieval Based on the Balanced Tree in Peer to Peer Systems*

Hai Jin and Jie Xu

Cluster and Grid Computing Lab
School of Computer Science and Technology
Huazhong University of Science and Technology, 430074, Wuhan, China
{hjin, jiexu}@hust.edu.cn

Abstract. Most existing Peer to Peer (P2P) systems support name-based retrieval and have provided very limited support for the full-text search of document contents. In this paper, we present a scheme (TRES-CORE) to support content-based retrieval. First, we propose a tree structure to organize data objects in vector-format in the P2P system, which is height-balanced so that the time complexity of search can be decreased. Second, we give a simple strategy for the placement of tree’s nodes, which can guarantee both load balancing and fault tolerance. Then an efficient policy for the query is given. Besides theoretical analysis that can prove the correctness of our scheme, a simulation-based study is carried out to evaluate its performance under various scenarios finally. In this study, it shows that using this content-based retrieval scheme (TRES-CORE) is more accurate and more efficient than some other schemes in the P2P system.

1 Introduction

Peer to Peer (P2P) systems have wide applications in many fields in recent years, such as file sharing, distributed computing and so on. Information retrieval is the key technology for file sharing. However, traditional approaches have either been centralized or used flooding to ensure the accuracy of results returned and most of them only provide name-based retrieval, that is, the user can not search a data object unless he knows its name. They lack support for content-based retrieval.

Current P2P retrieval technologies can be classified into three types. First, a centralized index is maintained at a server, and all queries are directed to this server. However, a centralized search engine is not suitable to be scalable, which can not perform the efficient retrieval in the P2P system and it is also a single point of failure, such as Napster [1]. Second, a distributed index is employed. The query will be flooded across the network to some other peers. But, network traffic

* This work is supported by National Science Foundation of China (NSFC) under grant No.60433040 and by China CNGI Projects under grant No.CNGI-04-12-2A, CNGI-04-12-1D.
generated by these flooding queries becomes un-scalable in large environments and it will lead to the poor network utilization. An example of this approach is the Gnutella system [2]. The third approach is the Distributed Hash Table (DHT) based scheme where the peer and the data object are structurally organized by a hash function. A query can get the result in $O(\log N)$ hops and it can generate fewer traffic in comparison with flooding-based mechanisms. Whereas, it can only support exact match queries and incurs the overhead maintaining the structure. Chord [3], Can [4], Pastry [5] and Tapestry [6] are examples of this approach.

In this paper, we explore the content-based retrieval scheme in P2P systems. First, traditional information retrieval techniques [7][8] are used to extract feature vectors from data objects. Using feature vectors of all data objects, a balanced search tree structure is formed. Then based on this search tree, we give an efficient retrieval scheme. And the time complexity of searching is $O(\log_B N)$ because the tree is height-balanced where $B$ is the balancing factor of the tree. Our simulation results show using our content-based retrieval scheme (TRES-CORE) can increase recall and reduce the network traffic, that is, it can improve the efficiency of query routing.

The rest of this paper is organized as follows. In section 2, we present related works to our work. Section 3 explains basic ideas of our information retrieval scheme. Section 4 discusses some improvements to the basic design in order to provide load balancing, fault tolerance and efficiency. Experimental results are presented in section 5, and the last section gives conclusions and future works of our work.

2 Related Work

There are also some of today’s works in the P2P information retrieval focusing on the content-based search. We describe them as follows.

A Hierarchical Summary Structure is proposed in [9], which employs three levels of summarization, naming as unit level, peer level and super level. However, in each level summary, how it is organized is not explained. And this is a key problem, which is able to result in the liner time complexity for the search if it is not organized well. Furthermore, it is another problem that how the feature vector of super peers and ordinary peers is generated accurately, which can effect the recall for the retrieval operation.

PlantP [10] presents a distributed content-based search algorithm in P2P systems. An inverted (word-to-document) index of the data objects that the peer wishes to share is created in each peer, and this index is summarized in a compact form. Then the summary is diffused throughout the network. Using these summaries, any peer can query and retrieve matching information from the collective information store of system. However, it is suitable for the multi-keyword-based retrieval but not for content-based retrieval using an inverted index.

The basic idea behind EZSearch [11] is in the following. Peers are partitioned into clusters. Each cluster contains peers having similar contents and manages a