Chapter 6
Scalable Index and Data Management for Unstructured Peer-To-Peer Networks

Shang-Feng Chiang  
National Taiwan University, Taiwan

Kuo Chiang  
National Taiwan University, Taiwan

Ruo-Jian Yu  
National Taiwan University, Taiwan

Sheng-De Wang  
National Taiwan University, Taiwan

ABSTRACT

In order to improve the scalability and reduce the traffic of Gnutella-like unstructured peer-to-peer networks, index caching and controlled flooding mechanisms had been an important research topic in recent years. In this chapter the authors will describe and present the current state of the art about index management schemes, interest groups and data clustering for unstructured peer-to-peer networks. Index caching mechanisms are an approach to reducing the traffic of keyword querying. However, the cached indices may incur redundant replications in the whole network, leading to the less efficient use of storage and the increase of traffic. They propose a multiplayer index management scheme that actively diffuses the indices in the network and groups indices according to their request rate. The peers of the group that have indices with higher request rate will be placed in layers that receive queries earlier. Their simulation shows that the proposed approach can keep a high success query rate as well as reduce the flooding size.

DOI: 10.4018/978-1-60566-661-7.ch006
INTRODUCTION

With the growth of the Internet, peer-to-peer (P2P) systems have become an important paradigm in designing large scale distribution systems. Peer-to-peer systems (Androutsellis-Theotokis, & Spinellis, 2004) provide effective ways of sharing data and can be based on overlay networks, which are classified by the degree of decentralization. The three categories are as follows: purely decentralized, partially decentralized, and hybrid decentralized architecture. Supporting efficient search of desired documents has been the most important issue in a decentralized peer-to-peer network. The overlay for decentralized peer-to-peer networks can be either unstructured or structured based on some distributed hash functions. Gnutella and Napster are pioneers in peer-to-peer file sharing systems and belong to the unstructured ones.

A class of structured peer-to-peer networks uses DHT (Distributed Hash Table) to maintain the shared documents. Distributed hash tables (DHTs) make use of hashing functions to provide distribution and lookup services. In this way, any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from names to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. DHTs are typically designed to scale a large number of nodes and handle nodes’ arrival and departure. With a routing table, all the participating nodes only need to communicate with a small fraction of all the nodes in a structured overlay network.

On the other hand, the unstructured peer-to-peer networks often rely on flooding mechanisms to search the desired objects. As a result, it needs to use techniques like index caching, active replication, or controlled flooding, to reduce the query traffic. The search algorithm of Gnutella use a kind of flooding method to discover objects, which sends queries to all nodes within a given TTL value. However, the mechanism is not scalable, since the query messages will grow exponentially due to its blind search method.

In this chapter, we will discuss some scalable techniques for index and data management for unstructured peer-to-peer networks. The concepts of interest group and data clustering will also be addressed. We also propose an index diffusion scheme to maintain a high success query rate and reduce the traffic load, for unstructured peer-to-peer systems.

BACKGROUND AND RELATED WORK

BitTorrent

BitTorrent is a peer-to-peer communication protocol (Cohen 2002) that can distribute large amounts of data widely without the original distributor incurring the entire costs of hardware, hosting, and bandwidth resources. Instead, when data is distributed using the BitTorrent protocol, each recipient supplies pieces of the data to newer recipients, reducing the cost and burden on any given individual source, providing redundancy against system problems, and reducing dependence on the original distributor.
Related Content

Towards Transparent Throughput Elasticity for IaaS Cloud Storage: Exploring the Benefits of Adaptive Block-Level Caching
[www.igi-global.com/article/towards-transparent-throughput-elasticity-for-iaas-cloud-storage/136251?camid=4v1a](www.igi-global.com/article/towards-transparent-throughput-elasticity-for-iaas-cloud-storage/136251?camid=4v1a)

Adaptable Information Provisioning in Collaborative Networks: An Object Modeling Framework and System Approach
[www.igi-global.com/article/adaptable-information-provisioning-collaborative-networks/58633?camid=4v1a](www.igi-global.com/article/adaptable-information-provisioning-collaborative-networks/58633?camid=4v1a)

Micro-Services: A Service-Oriented Paradigm for Scalable, Distributed Data Management
[www.igi-global.com/chapter/micro-services-service-oriented-paradigm/62822?camid=4v1a](www.igi-global.com/chapter/micro-services-service-oriented-paradigm/62822?camid=4v1a)

MaGate: An Interoperable, Decentralized and Modular High-Level Grid Scheduler
[www.igi-global.com/article/magate-interoperable-decentralized-modular-high/46048?camid=4v1a](www.igi-global.com/article/magate-interoperable-decentralized-modular-high/46048?camid=4v1a)