A Study on Hierarchical Peer-to-Peer Systems to Realize Real-Time File Search
(実時間ファイル検索のための階層型ピア・ツー・ピアシステムの研究)

Recently, Peer-to-Peer (P2P) systems have attracted considerable attention as a way of
providing scalable network services over the Internet. A P2P system consists of a large number of
computers called peers connected with a logical network called P2P overlay, each peer participating
in a P2P can simultaneously play the role of a server and a client, and services such as file sharing
and content delivery are provided among peers in a peer-to-peer manner. A key challenge in P2P
systems is how to quickly identify the location of a target file existing in the network since they do not
rely on a centralized server as in C/S systems. In the existed literatures, the use of hierarchical
structure is currently being the most promising approach to realize a quick access to such location
information with a reasonably low cost. Unfortunately however, it has not yet been clarified that how
we can realize a real-time search of update location information in hierarchical P2P systems, since it
lacks a mechanism of reflecting dynamic change of location information to the search result. Thus we
proposed a new hierarchical model focusing on this issue. We also introduced the algorithms and
schemes developed for improving the searching efficiency and accuracy in P2P file sharing systems.
Following are the content of this thesis.

In chapter 1, we introduced the background and related works of P2P systems. It includes
existing file search techniques both in structured and unstructured P2P systems, social tagging
approach used in P2P tagging system, as well as popular caching policies adopted in P2P systems
to reduce the heavy network load. To overcome the limitations of conventional works, we proposed
the new hierarchical architecture.

In chapter 2, we described the new hierarchical P2P file sharing architecture in order to realize a
quick search of update location information in distributed P2P environment. The hierarchical
architecture consists of three layers; i.e., the correlation between files (held by user peers in the
bottom layer) and peers in the middle layer is controlled by the peers in the top layer (in this
architecture, a peer in the top layer is referred to as a central server and a peer in the middle layer is
referred to as a sub-server). We implemented a prototype of the proposed system, and conducted
several experiments with one central server, 100 sub-servers and 1000 user peers. The results of
experiments indicate that it completes an upload of 10000 file indexes to relevant sub-servers in a
few minutes and a query forwarding to a relevant peer within 100 ms.

In chapter 3, we introduced a file search algorithm for hierarchical P2P file search system. The
proposed scheme consists of two parts. The first part determines a way of associating files held by
each peer in the bottom layer to sub-servers in the middle layer, where each sub-server plays the
same role with the (centralized) server in conventional search engines. The second part provides a
way of forwarding a query received by the central server in the top layer to an appropriate sub-server
relevant to the query. The proposed scheme is based on the notion of tags, and a technique of
priority sequence of tags is introduced to the central server, in order to realize a quick forwarding of
received queries. The result of preliminary performance evaluation indicates that the number of tags which must be examined in forwarding a given query is bounded by a small constant.

In chapter 4, we described an automatic cluster-based tag attachment scheme to improve the quality of file search in P2P environment. The proposed scheme combines text clustering with a modified tag extraction algorithm, and is executed in a fully distributed manner. Meanwhile, the optimal cluster number can also be fixed automatically through a distance cost function. The result of experiments indicates that the proposed approach is capable of making effective and efficient tag attachment in real scenarios; i.e., for more than 90% of documents, it attaches the same tags as the ones attached by human reviewers. Moreover, it proofs by the experiments that the optimal cluster number is almost the same as the number of topics from the website.

In chapter 5, a two-level caching protocol for hierarchical P2P file sharing systems is proposed to relax bottlenecks and heavy workload in such systems. The simulation result indicates that our caching protocol reduces the network traffic by 52% and exhibits a higher hit rate even in small cache sizes.

In chapter 6, we made a conclusion of this thesis.