A System Implementation of Gathering Web-Links on Distributed Computing Environment

Masanori Ito

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2002

Keywords: Hyperlink, Distributed Computing, World Wide Web, ORB

Modern society, WWW has affect human-life as one of the media. The features of WWW are Large-scale, quick change, existing various contents and structural. Recently, research and system of Web-Links Structural are proposed. Generally, WWW and Web-links are collected by large-scale distribution parallel WWW Robots. But this method also can't collect all WWW.

Therefore, many researchers changes research-point and research the approach to arrive at the destination quicker in collection-scale. Distributed systems for search WWW exist, but research investigated how to construct or technical problem is not found.

The purpose of this research is the implementation of Gathering Web-Links on Distributed Computing Environment, and to clarify technical problem acquired in implementation process.

First, this paper describes the points of design and function of gathering Web-Links system, and explain outline of hardware, distributed system technology, programming language and database.

Next, this describes flow of processing of the system and verified whether system would implement as design. This system extracts hyperlink from Web page on Internet. This extracting process is better by distributed parallel computing. Gathering Web-Links system require database to store large link data. This system is distributed computing on HORB. This system consist of 10 Slave PC, 1 Master PC, 1 Database Server. Slave PC is Gathering Web-Links. Master PC manages tasks of Slave PC. Programming language is Java, this have

multi-thread and asynchronous method.

Next, this paper describes two-evaluation experiment of system. 1st experiment investigated influence of time and number of Slave PC to system performance. 2nd experiment investigated influence of tasks to system performance. Tasks are URL data set and link depth.1st showed performance per Slave PC set is decreasing with time. This result is to overlap same Web pages.2nd showed performance per Slave PC set is increasing with amount of tasks. This result is to decrease overlapping. But amount of tasks is not necessarily good. Amount of tasks increase overlap of Web pages or connectivity of Web links with time. And these experiments showed to influence network traffic to system performance.

Finally, this paper describes problem and improving idea of system based on experiment. These problems are connection to no-react Server, processing time of database, un-corresponding to CSS. Improvement ideas are method of distributed processing and distributed database system.