December 2003

Using P2P Computing and Mobile Agents For Web Information Retrieval: A Framework Design

Yihua Sheng
Southern Illinois University, Carbondale

USING P2P COMPUTING AND MOBILE AGENTS FOR WEB INFORMATION RETRIEVAL: A FRAMEWORK DESIGN

Yihua Philip Sheng
Southern Illinois University, Carbondale
pshan@cs.siu.edu

Abstract

The Internet has tremendous amount of information to be readily gathered and mined. But lack of useful tools for gathering and mining information makes the information under-utilized. Also, it would be very nice if an individual could have his own information gathering and filtering system so that he can freely tweak information according to his own needs. But building and maintaining a reliable and efficient web information retrieval system (WIRS) is associated with a very high cost; reliable and durable software and hardware, a large number of computers for web crawling, and high speed network connection are the minimum requirements. This paper proposes a lightweight information retrieval system based on mobile agents and peer-to-peer (P2P) computing technologies to drastically reduce the cost.

P2P computing is a new networking paradigm where two or more clients work together as equals. Harnessing the unused processing power and storage of computers on the Internet could deliver supercomputing capabilities at a fraction of the current cost. The SETI@home project and Naspter are two successful stories. A mobile agent is a persistent entity (i.e. it can outlive the application it originates from) which is typically limited in size and most importantly is able to migrate, i.e. to suspend its execution, move to another location, and continue execution there. When a mobile agent migrates, it has to carry its data state (variables) as well as execution state (active threads).

An execution environment for mobile agents, called an agency, will be voluntarily installed on each participating user's computer in the P2P computing pool. The function of the agency is to accept, run, and move mobile agents. Four different kinds of mobile agents are present in this system: Web Crawling Agent (WCA), Information Filtering Agent (IFA), Manager Agent (MA), and Report Generation Agent (RCA). WCAs are responsible for retrieving information from different sources on the Internet. The input to a WCA is a list of Internet URL addresses needed to crawl. The output of a WCA is the contents of web pages crawled. IFAs are responsible for filtering information received from WCA. The precision of the information returned largely relies on IFA. An IFA decides if the information from WCAs should be discarded or sent back as useful information. A single IFA can receive and filter information from multiple WCAs. MAs are responsible for managing mobile agents (WCAs and IFAs) in the system, including themselves. Each MA maintains a reasonable number of other agents in the system. RCAs are responsible for generating final reports based on the returned web pages stored in the database. Report templates are created and stored in a Report Templates Database. Users can define their own templates and store them in the database. To some extent, templates guide the activities of web crawling and information filtering. In addition to these mobile agents, a stationary console agent (CA) is also developed and installed with the agency in users' computers. The coordination and collaboration among these agents are described below:

1. *A user uses the CA to start an information retrieval job, such as traversing a set of websites and their related sites to extract information on a particular product.*

2. *After receiving a search job from a user, the CA analyzes the nature of the job and then splits it into smaller pieces. MAs are created for each piece of the job and dispatched to the P2P computing pool for execution.*

3. *MAs then create a series of WCAs and IFAs and send them out to the computers in the same pool.*

4. *The WCAs crawl websites and send the web pages to their corresponding IFAs.*

5. *The IFAs parse received web pages and then either discard them or send them needed content to a designated database. The database resides on a cluster of computers in the same P2P computing pool.*

6. *Each MA is responsible for the WCAs and IFAs it creates. A MA monitors the status and progress of its WCAs and IFAs and reports to the user through the CA. If necessary, a MA can also split the job further by creating a number of sub-MAs and send them out to other computers to speed up the whole process.*

If the system load of a computer on which a mobile agent is executing exceeds preset thresholds, the agent will move to another available computer in the P2P pool and notify the change to its corresponding agents. WCAs might also need to move to different computers in the P2P pool due to a detected blockade by the destination system from which information is being gathered.