Design of Internet Search Engine by Intelligent Agents on WWW

Ryota NAKANO Masato NOTO

Department of Electrical Engineering, Kanagawa University, Yokohama, 221-8686 JAPAN
Phone: +81-45-481-5661 (Ext. 3216) Fax: +81-45-491-7915
E-mail: ryouta@cc4-4.kanagawa-u.ac.jp noto@cc.kanagawa-u.ac.jp

Abstract: The Internet has become widely used in many countries. In particular, a new emerging technology, the WWW (World Wide Web), which has become a major application of the Internet, has been rapidly developed. As a result, there are hundreds of millions of URLs (Uniform Resource Locators) on the WWW, and the total number of URLs is still explosively increasing. To get information from the WWW, we generally use Internet search engines. However, we cannot always get the actual information we want. Accordingly, we have solved this problem by constructing a prototype system based on agents by programming language Java for constructing a more effective search engine.

This so-called "intelligent agent system on WWW" deletes redundant HTML (Hyper Text Markup Language) files and exchanges information about the existence of URLs. And we found that our prototype system is more powerful and effective than conventional search engines.

1 Introduction

The Internet has become a major topic in newspapers, magazines, and various media, such as a television, over the last several years.

The Internet influences not only the world of computers but also various fields such as business and education. In 1995, the number of countries connected to the Internet was more than 150, and it is said that the number of users is now more than forty million. The Internet has spread rapidly because it:

- uses a TCP/IP protocol attached to UNIX,
- has an open protocol specification based on the C language,
- has high speed and flexibility,
- uses a standard protocol such as that used by LANs like Ethernets.

The Internet is a world network even though it began as an object of research rather than business. It originally mainly provided high-bandwidth connectivity between major computing sites in government, educational, and research laboratories. And the origin of applications such as electronic mail and the WWW (World Wide Web), a new information system which is available on the Internet, is the tool developed by researchers (such as staff members and students) for person-to-person communication.

The structure of WWW is shown in Figure 1.

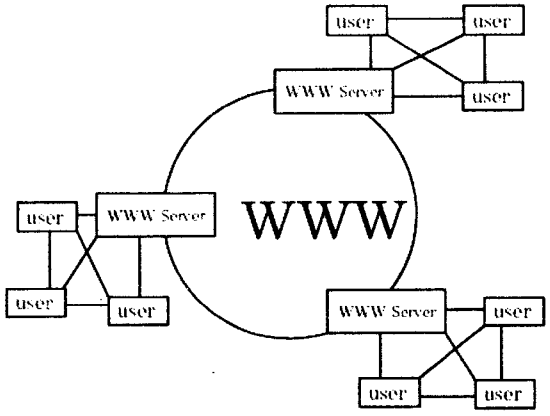

Figure 1: Structure of WWW

The Internet has recently been widely used in many countries. In particular, the WWW, which has become a major application of the Internet, has been rapidly developed. As a result, there are hundreds of millions URLs (Uniform Resource Locators) on WWW sites all over the world; the total number of URLs is still explosively increasing. To get information from the WWW, there are a lot of search engines. And in the following section the two kinds of search engine are outlined.

2 Search Engine

Search engines are generally used to get information from the WWW. The characteristics of the two types of search engine, categorical search and robot search, are explained below.

2.1 Categorical Search Engine

"Yahoo[1]" is one of the most popular categorical search engines. The administrator of the search engine (for example, Yahoo) register URLs and contents in the search engine and, thus, make a database. Then, a person inputs a few keywords into the search engine and quickly receives the results of the search. It is difficult to artificially divide a category because of the increasing amount of information on the WWW.

2.2 Robot Type Search Engine

"Fast search[2]" is the one of the most popular robot-type search engines. Software called a *robot* accesses a WWW server, and extracts a key word from the HTML (Hyper Text Markup Language) file contained in the WWW server. When it accesses a certain WWW server, a hyper-link is further used, and the file of another WWW server is referred to. This procedure is repeated and information is collected. The structure of the robot reference is shown in Figure 2.

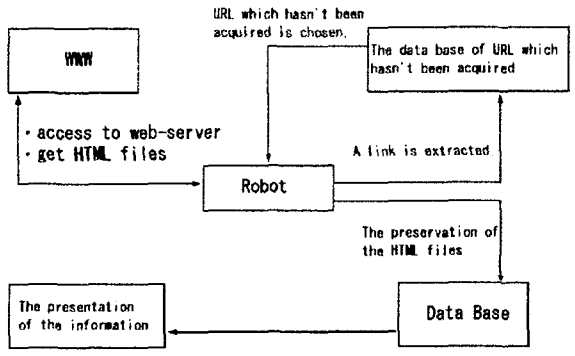

Figure 2: Robot Type Search Engine

We usually get many more answers from an automatic robot type search engine than from a categorical one. However, we cannot always get the actual information we want. Robot type search engines suffer the following problems:

- URLs denoted by robot type search does not exist
- Representation of redundant URLs
- Deterioration of information with time

These problems can be solved by using our developed agent.

3 Intelligent Agent

3.1 Agent

The technical term *agent* has been used in many fields of research; its definition is therefore ambiguous[3]. However, it is said that an agent can solve its own problems. And an agent can also exchange information with other agents so that it can accomplish various tasks. They generally have the following four characteristics:

- Autonomy: Agents act without direct interference from humans and other systems. They can therefore control their own behavior and conditions.
- Social ability: Agents are capable of interacting with other agents and humans in order to satisfy their design objectives.
- Reactivity: Agents are able to perceive their environment and respond in a timely fashion to changes that occur in it in order to satisfy their design objectives.
- Pro-activeness: Agents are able to exhibit goal-directed behavior by taking the initiative in order to satisfy their design objectives.

A rough outline of an agent system is shown in Figure 3.

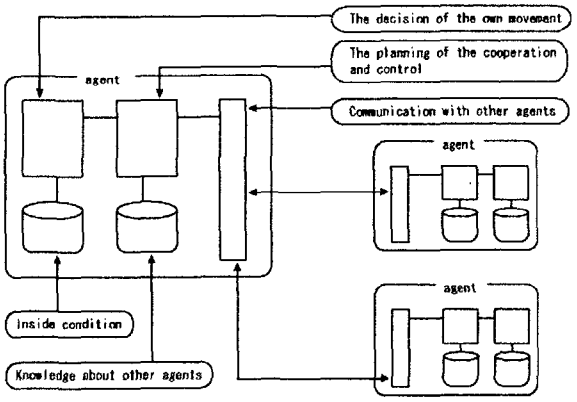

Figure 3: The Agent System

3.2 Intelligent Agent

An agent is a computer system situated in an environment, and it is capable of autonomous action in this

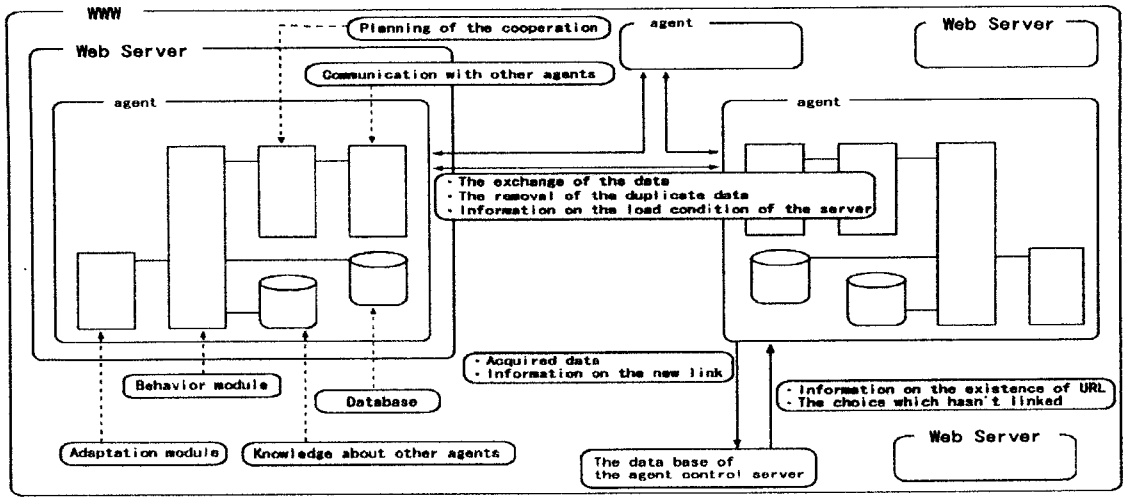

Figure 4: The System of Intelligent Agent

environment in order to meet its design objective. In addition, we define *intelligent agent* by extending the conventional agent definition. Its characteristics are that it can solve its own problem by itself, exchange information between other agents, accomplish tasks, and actively move in the environment[4].

We have developed an autonomous and cooperative intelligent agent for deleting redundant HTML files and for exchanging information about the existence of URLs. And we use this agent in a prototype system that is more powerful and effective than conventional search engines. This intelligent agent has the following special functions:

- Acting: process HTML files without using browser instead of human being.
- Spontaneity: satisfies a user's requirement by itself.
- Adaptability: can be efficiently decided the source of information that can access in dynamic on WWW changing every moment.
- Cooperation: mutually convey the redundancy of the information and the load conditions of the server by cooperation between agents.

The agent functions as follows:

1. runs on the WWW and acquires HTML files
2. exchanges information by cooperation with other agents
3. acquires HTML files and stores them in a file system (database)
4. deletes duplicated URLs
5. regularly confirms existence of the stored data
6. sends the requested information to the user

Figure 4 shows an outline of the developed agent system.

Our prototype system has the following functions:

- Behavior module: The module that planning of self and to execute according to an environment module. And it gets HTML files and stores them in the database, decide how to patrol on WWW.
- Adaptation module: The module that it moves in the background of the above module (Behavior module) and which the optimization of the own movement is attempted to corresponding to the outside environment.
- Communication with other agents: Actually, the part which communicates with the agent.
- Planning in order to cooperate with other agents.
- Knowledge about other agents: plans the above cooperation efficiently.
- Database: stores data of HTML files, un-link and URL information.

4 Simulation

Information on the WWW is acquired by using the programming language Java. We used FreeBSD 3.4

as the operating system (OS). And we used JDK1-1-8 (javac) for the compiler[5].

4.1 Acquisition of URLs

The resources on the WWW are distinguished by a URL, which has a form as follows:

URL

```
protocol://host:port/file
```

As for the host parameter with ftp and http, the machine of the place of the acquisition is distinguished with protocol used frequently resource. As for the port parameter of the option designation, the software port of the server is distinguished. When this parameter is omitted, a fixed value is used in as the protocol File. As for the parameter, a file on the server is distinguished. The result of programming of acquiring resource of URL of the WWW top in the following is shown.

The executive result of the **HttpClient.java** program acquires the html file of the URL in order to connect with the server which is specified. GetURL connects **HttpClient.java** with the Web server directly and communicates by using a HTTP protocol though GetURL according to the URL class for the details of the protocol treatment. When an executive result is seen, it is connected with the Web server, and it is understood that a file is acquired. Because it is a Web server, it thinks that an agent actually gets around this program to become a basis in this system.

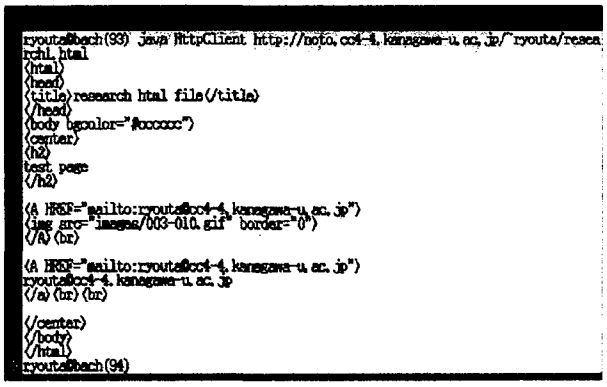

Figure 5: The result of **HttpClient.java**

4.2 Download of Information on URL

The **GetURLInfo.java** program is the class which indicates provides information about the URL.

Information such as contents, size, and final update day about the resource which it is being referred to by the URL is acquired by using URLConnection.

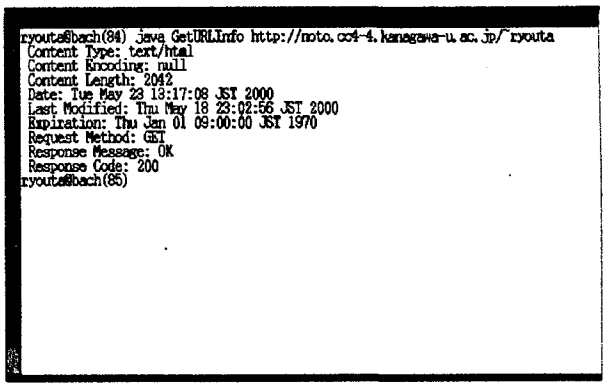

Figure 6: The result of **GetURLInfo.java**

We have understood that information at the URL can be acquired when this program is run. When an agent acquires an HTML file, this program records data such as a final update day.

5 Conclusion

We have developed the prototype of an Internet search engine that improves the conventional engine by using an agent on the WWW. We will soon complete the search engine on WWW. To make more powerful search engine, we must solve the following problems:

- Introducing an active learning agent,
- Techniques and methods to get a large number of data,
- Flexible human interfaces.

Acknowledgement

This work is partially supported by the Grant from Ministry of Education, Science and Culture of Japan, No. 12780246.

References

[1] http://www.yahoo.com/

[2] http://www.ussc.alltheweb.com/

[3] Fah-Chun Cheong. "Internet Agent", 1996

[4] Gerhard Weiss. "Multiagent Systems", 1998

[5] David Flanagan. "Java in a Nutshell", 1997