(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 December 2000 (07.12.2000)

PCT

(10) International Publication Number
WO 00/74294 A2

(51) International Patent Classification[7]: H04L

(21) International Application Number: PCT/KR99/00356

(22) International Filing Date: 3 July 1999 (03.07.1999)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1999/19679 31 May 1999 (31.05.1999) KR

(71) Applicant *(for all designated States except US)*: **WEBNARA CO., LTD.** [KR/KR]; Technomart 29F-10, 546-4, Guui-dong, Kwangjin-gu, Seoul 143-200 (KR).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **KIM, So, Young** [KR/KR]; #92-201, Siyoung Apt., Jangan-2dong, Dongdaemoon-gu, Seoul 130-102 (KR).

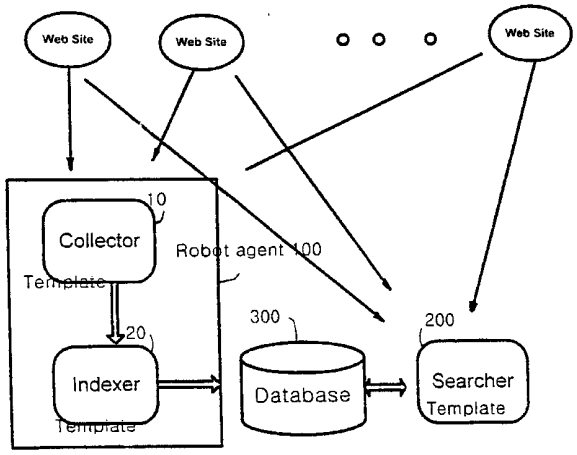
(81) Designated States *(national)*: AL, AM, AU, BA, BB, BG, BR, CA, CN, CU, CZ, EE, HR, HU, ID, IL, IN, IS, JP, LC, LK, LR, LT, LV, MG, MK, MN, MX, NO, NZ, PL, RO, SG, SI, SK, SL, TR, TT, UA, US, UZ, VN, YU.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: GENERAL-PURPOSE ROBOT AGENT AND REAL-TIME SEARCH METHOD



(57) Abstract: The present invention provides a general-purpose robot agent (100) and real-time search method for freely operating individual modules as independent systems or an integrated system and allowing each module to be easily integrated into another system as a component by creating templates for individual modules such as collector (10), indexer (20), and searcher (200), which were coded within the program in the conventional techniques, to construct the modules in an independent format, thereby increasing flexibility of the system. The present invention also provides a general-purpose robot agent (100) and real-time search method which employs an indexing method depending upon a common format instead of specific terms. Moreover, the present invention provides a general-purpose robot agent (100) and real-time search method, where a searcher (200) not only searches a database (300) but also directly searches predetermined web sites, thereby effectively searching contents on the web sites having short update cycles.

WO 00/74294 A2

GENERAL-PURPOSE ROBOT AGENT AND REAL-TIME SEARCH METHOD

Technical Field

The present invention relates to a web search engine and more particularly to general-purpose robot agent and real-time search method.

Background Art

Typical web search engines (or search sites) serviced in present are based on a robot agent system usually called a robot agent, a database, and a search step.

Referring to FIG. 1, a conventional web search engine program (or web search system) will be described in detail.

The web search system comprises a robot agent system 100, a search system (hereinafter, called a searcher) 200, and a database 300. The robot agent system 100 comprises a collecting system (hereinafter, called a collector) 10 and an indexing system (hereinafter, called an indexer) 20.

In the robot agent system 100, the collector 10 collects data in predetermined web sites and the indexer 20 indexes the data by analyzing morphemes and processing stop words and particles under the linkage with electronic dictionaries including dictionaries of nouns, synonyms, and particles.

The indexed data is stored in the database 300, and thereafter, the searcher 200 searches the stored data.

To search information using the web search system usually means to search the database 300 using the searcher 200.

However, the conventional internet search engines have the following problems.

Primarily, since existing web search systems developed with particular purposes have the collector 10 and the indexer 20 coded as parts of the entire program, the collector 10 and the indexer 20 are strictly subjected to the entire program.

This decreases flexibility of the systems in extension of their functions for other purposes or switch to systems of another use. Furthermore, administration and maintenance of the systems costs a great deal since technical knowledge of software is required.

For example, a web search system produced for the purpose of searching shopping malls on the web comprises a collector for collecting only shopping mall web sites and an indexer for indexing data related to the collected shopping malls and constructs a program in linkage with an electronic dictionary related to the shopping mall.

However, if this web search system is used for any other purpose other than the shopping mall, the collector 10 and the indexer 20 which have been coded for the shopping mall and the searcher 200 should be all altered.

Such alteration requires technical knowledge of software, thereby making free and convenient alteration of the search system difficult.

The conventional indexing method comprises the steps of analyzing morphemes and processing stop words and particles. However, these steps result in difficulty in constituting the electronic dictionary. Moreover, the electronic

dictionary itself changes according to service fields it is applied to, and interpreting and processing inevitably results in hard coding within the program, thereby decreasing flexibility and extensibility of the entire system.

The conventional searcher 200 searches the contents of the database 300. In this structure, if updated contents of the web sites have not been reflected in the database 300 yet, the searcher 200 retrieves old contents excluding updated new contents.

Consequently, the conventional robot agent has a problem that it cannot search updated new data if the data on the web sites is frequently changed.

Summary of the Invention

To overcome the defects described above, the present invention provides a general-purpose robot agent and real-time search method for freely operating individual modules as independent systems or an integrated system and allowing each module to be easily integrated into another system as a component by creating templates for individual modules such as collector, indexer, and searcher, which were coded within the program in the conventional techniques, to construct the modules in an independent format, thereby increasing flexibility of the system.

The present invention also provides a general-purpose robot agent and real-time search method which employs an indexing method depending upon a common format instead of specific terms.

Moreover, the present invention provides a general-purpose robot agent and real-time search method, where a searcher not only searches a database but also directly searches predetermined web sites, thereby effectively searching contents on the web sites having short update cycles.

Brief Description of Drawings

The present invention will become more fully understood from the detailed description given hereinbelow and the accompanying drawings which are given by way of illustration only, and thus are not limitative of the present invention, and wherein:

FIG. 1 is a schematic diagram of a conventional web search system; and

FIG. 2 is a schematic diagram of a general-purpose robot agent and real-time search method according to the present invention.

Best Mode for carrying Out the Invention

The present invention relates to a general-purpose robot agent and real-time search method and is developed based upon Pure Java, thus being operated in most existing platforms without a special porting process.

A robot agent 100 comprises a collector 10 for performing a collecting process and an indexer 20 for performing an indexing process.

Data indexed by the indexer 20 in the indexing process constructs a database 300. In a search process, the database

300 is searched by a searcher 200.

During the actuation, or the initialization of the web search system, a processor is controlled by an initialization file provided from the outside. The initialization file is a simple text file which can be easily edited according to rules. The system is controlled by this file.

The initialization file for the collector 10 contains initial URLs of sites, the number of threads, and a collection/removal pattern.

The initialization file for the indexer 20 contains index rules of respective sites. The initialization file for the searcher 200 contains collected data/index rule and a display format for each site.

In conventional techniques, control information of the collector 10, indexer 20 and searcher 200 is coded within the program. On the other hand, the present invention creates templates for the control information of the collector 10, indexer 20 and searcher 200 in the outside of the system. The functions located outside the system can be easily changed in accordance with different purposes, thus being used as a general-purpose robot agent. Each sub-system is also designed to function as an independent system, so it can be freely utilized for different particular purposes.

The conventional indexing method uses specific phrases or words with respect to the web file. On the other hand, the present invention indexes data of the web file based upon a predetermined pattern shown in the web file, so the conventional morpheme analysis and stop words and particle

processing are not required, thereby decreasing difficulties in construction of an electronic dictionary and change of the electronic dictionary itself according to which fields the dictionary is applied to. Consequently, the present invention improves flexibility and extensibility of the entire system a lot.

In the search process, the searcher 200 not only searches the database 300 but also directly searches specified web sites, so the web sites whose updated new data cannot be searched by the conventional methods can be searched.

Through such operations, the general-purpose robot agent and real-time search method of the present invention can be generally applied to general search systems, special search systems for shopping malls, special search systems for book information, and so on.

The operation of each module in the general-purpose robot agent having such features will be described.

The collector 10, which is a sub system of the robot agent, basically performs a web server mirroring function as in existing other search systems.

The collector 10 of the present invention is different from the conventional one in that it can ensure the flexibility through a template function. Specifically, a collection skip URL pattern can be sophisticatedly defined or files to be collected can be limited in associated with the indexing to achieve efficient collection control. In addition, taking network traffic into consideration, a thread can be assigned to each site.

The indexer 20 is basically controlled by the template, namely, the initialization file.

Differently from the conventional indexing system, the indexer 20 of the present invention does not use electronic dictionaries and program routines subjected to service while analyzing or parsing files and constructing a database. Since the present invention indexes all files based upon only index rules recorded in the template, thus ensuring generality and achieving a high index rate through rapid index speed.

Index rule mechanism is based upon the idea that a web file to be indexed has a particular pattern for record of information.

For example, for the file of web shopping mall, a proper record pattern instead of different special patterns is repeatedly applied to individual articles when recording article information (such as name, price, and manufacturer of the article). Particularly, in case of providing mass information, it is general to use a specified proper pattern. In other words, in case of indexing data of the shopping mall, once the repeated pattern is written as an index rule according to defined grammar, the searcher receives this index rule and files and performs the indexing.

If a plurality of index rules are written in the same site, the searcher automatically chooses an index rule corresponding to the style of a file. The index rule is supported by a function of indexing only files containing a particular word, a function of processing extracted data

(e.g., removal and replacement of a particular character string), and a function of making information into a database (e.g., embedding a monetary unit or URL of a homepage in a certain field).

The searcher 200 is written in Java Servlet and divided into a meta search system and a directory (subject) search system in functional aspect. Since an idea of template is applied to the searcher 200 in the same way as to the robot agent 100, supplement of search sites or change of user interface elements can be achieved only by changing the initialization file. This makes the present invention overcome the defect of the existing searcher that should repeat modification of program, compile, and deployment even in case of trivial modification of information.

To overcome stateless (basically, the HTTP protocol does not define a session) which is one of serious obstacles in web application programs and to achieve an optimal response time which is essential in a method for searching web sites having short update cycles in real time, the following elementary techniques are used.

In the initialization file is described control information for most efficiently searching, namely, collecting and indexing search sites. The control information includes cache time, type of information (associated with a directory), shortest search path, and index rule (subset of the index rule of the indexer).

At the initial stage, generation and initialization of spare classes, generation of threads, and assignment of shared memory are performed to minimize an overhead

occurring during the service, using resource management.

An internal cache is implemented to take advantages of a local database. The cache comprises a memory cache and a disc cache and is embodied in the unit of keywords and the unit of subjects (directory). Specifically, when there is a request for service related to the same keywords and directories, the content in the cache is used for response. A cache manager is operated as a separate thread and automatically performs setting, removal, and conversion of cache.

Through such operation, the general-purpose robot agent and real-time search method of the present invention allows search engines to be applied to general search systems, search systems specialized in shopping malls, search systems specialized in book information, and other like systems.

In addition, the searcher of the present invention not only searching the database but also directly searches predetermined web sites having short update cycles in real time.

Although the preferred embodiments of the present invention have been disclosed for illustrative purposes, those skilled in the art will appreciate that various modifications, additions and substitutions are possible, without departing from the scope and spirit of the invention as recited in the accompanying claims.

WHAT IS CLAIMED IS:

1. In a web search system for searching information over the internet, a general-purpose robot agent and real-time search method comprising the steps of:

collecting data from predetermined web sites using a collector;

indexing the collected data using an indexer;

making the indexed data into a database; and

searching the data in the database using a searcher.

2. The general-purpose robot agent and real-time search method as claimed in claim 1, wherein, in said collecting, indexing, and searching steps, subordinate control information of a service is embodied in the form of templates.

3. The general-purpose robot agent and real-time search method as claimed in claim 1, wherein said indexing step is characterized by indexing the collected data based upon a form of description of a file instead of a content of the file.

4. The general-purpose robot agent and real-time search method as claimed in claim 1, wherein said searching step is characterized by not only searching the data in the database but also directly searching the web sites.

5. The general-purpose robot agent and real-time search method as claimed in claim 1, wherein said web search

system is developed based upon Java.

FIG.1

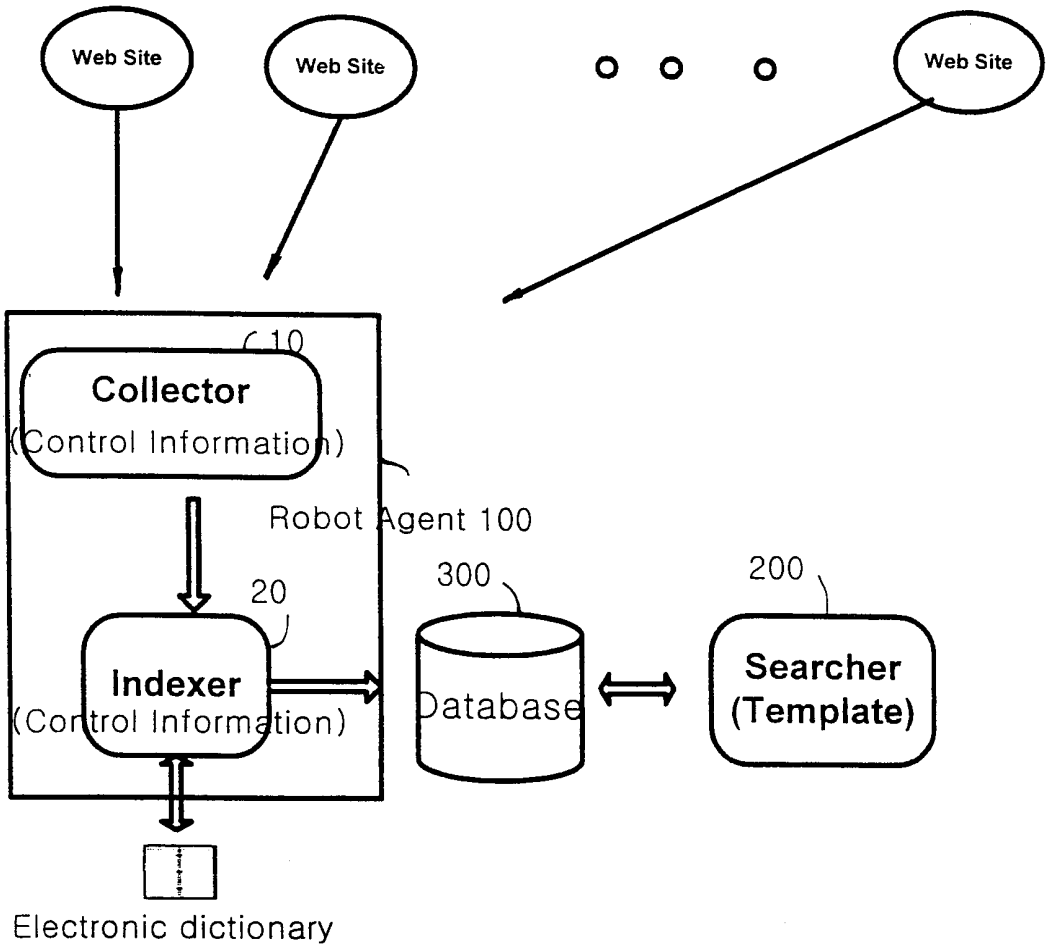
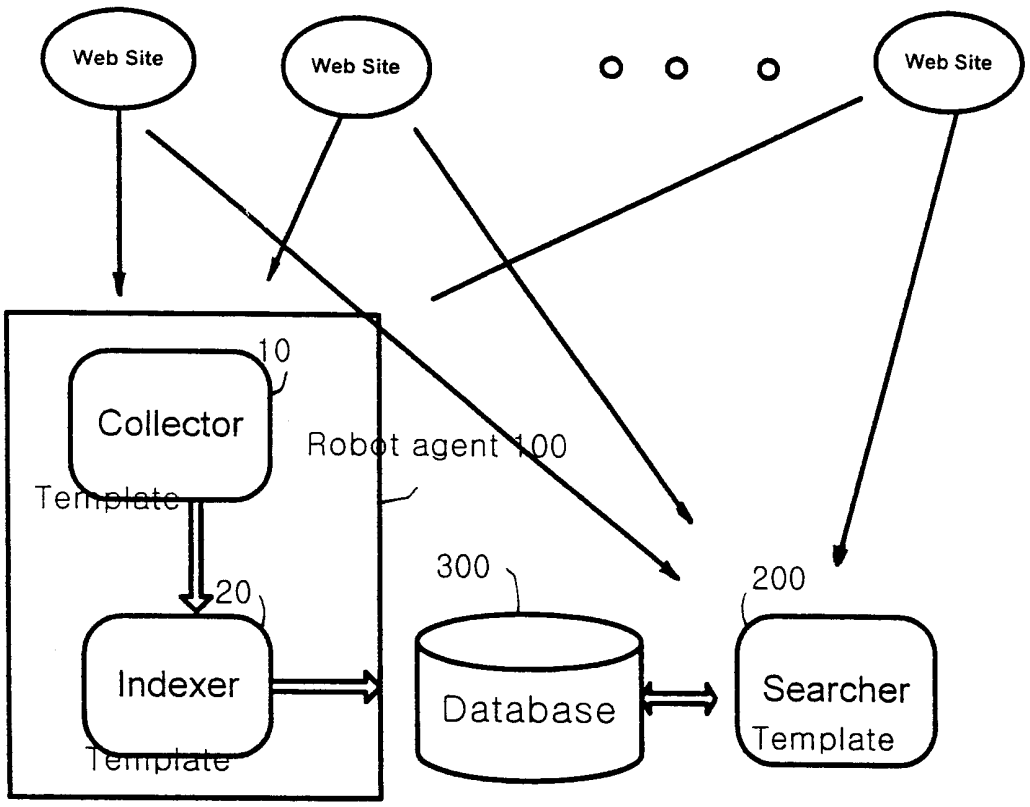

Web Site
Web Site
Web Site
10
Collector
(Control Information)
Robot Agent 100
20
Indexer
(Control Information)
300
Database
200
Searcher
(Template)
Electronic dictionary

FIG.2


Web Site
Web Site
Web Site
10
Collector
Robot agent 100
Template
20
Indexer
Template
300
Database
200
Searcher
Template