DESCRIPTION

[Title of Invention] RECORDING MEDIUM, DATA ACQUISITION METHOD, AND DATA ACQUISITION DEVICE

[Technical Field]

[0001] The present invention relates to a data acquisition program, a data acquisition method, and a data acquisition device.

[Background Art]

[0002] A crawler tool is known as a tool for collecting information disclosed on the Internet. The crawler tool visits websites on the Internet, and stores the content per uniform resource locator (URL), in other words, per page. Moreover, a method of providing information to a user has been developed in which the information is analyzed with ontology with a relation between lexicons defined from the page content that has been stored; useful information is cut out and is automatically extracted based on the rule; and pieces of information having the same meaning and any other piece of information are integrated.

[0003] Furthermore, a method of extracting information in a tabular format or a database format based on a tag included in a document in HyperText Markup Language (HTML), and extracting information having a relation with an attribute that is dependent on an item, and the attribute value, based on dependency between pieces of information indicated by the tag, has been developed. In a method of extracting pieces of information from the tabular format, a piece of information that is placed immediately above, immediately below, or immediately left side of information in the tabular format, is extracted as the item. A piece of information placed in the upper end row or the left end column in the tabular format is the attribute, and a piece of information placed below or on the right side of the

above is extracted as the attribute value.

[Citation List]

[Non Patent Citation]

[0004] Patent Document 1: Japanese Laid-open Patent Publication No. 2004-62446

Patent Document 2: Japanese Laid-open Patent Publication No. 2010-15202

[Summary of Invention]

[Technical Problem]

[0005] However, for example, when the ontology is used, a relation between lexicons needs to be registered in advance. Thus, it is difficult to extract information not included in the ontology. Moreover, for example, when predetermined positions in the tabular format are extracted as the item, the attribute, and the attribute value, when an item appears in the middle of the tabular format, the item is sometimes extracted as an attribute value. Hence, it is difficult to extract information corresponding to tables in various formats.

[0006] In one aspect, the present invention provides a data acquisition program, a data acquisition method, and a data acquisition device that can support the registration of a location of data to be extracted from an HTML document.

Solution to Problem

[Solution to Problem]

[0007] In one aspect, the computer-readable recording medium stores a data acquisition program that causes a computer to execute a process of extracting first data and second data that are separated from each other by a predetermined tag or symbol, from a document written in HTML. Moreover, the computer-readable recording medium stores the data acquisition program that causes the computer to execute a process of displaying the extracted first data and the extracted second data in an associated manner. Furthermore, the computer-readable recording

medium stores the data acquisition program that causes the computer to execute a process of specifying a position of a tag included in the document on a hierarchical structure with respect to the first data or the second data in the document, upon detecting that the displayed first data or the displayed second data is selected. Still furthermore, the computer-readable recording medium stores the data acquisition program that causes the computer to execute a process of allowing registration of the position on the hierarchical structure.

[Advantageous Effects of Invention]

[0008] It is possible to support the registration of a location of data to be extracted from an HTML document.

[Brief Description of Drawings]

[0009] FIG. 1 is a block diagram illustrating a structural example of a data acquisition device.

FIG. 2 is a diagram illustrating an example of a target storage unit.

FIG. 3 is a diagram illustrating an example of a page storage unit.

FIG. 4 is a diagram illustrating an example of an extracted data storage unit.

FIG. 5 is a diagram illustrating an example of a position storage unit.

FIG. 6 is a diagram illustrating an example of an HTML document and a selection screen of extracted data.

FIG. 7 is a diagram illustrating another example of an HTML document and a selection screen of extracted data.

FIG. 8 is a diagram illustrating another example of an HTML document and a selection screen of extracted data.

FIG. 9 is a diagram illustrating another example of an HTML document and a selection screen of extracted data.

FIG. 10 is a flowchart illustrating an example of a registration process.

FIG. 11 is a diagram illustrating an example of a

CLAIMS

1. A computer-readable recording medium storing a data acquisition program that causes a computer to execute a process comprising:

extracting first data and second data that are separated from each other by a predetermined tag or symbol, from a document written in HTML;

displaying the extracted first data and the extracted second data in an associated manner; and

specifying a position of a tag included in the document on a hierarchical structure with respect to the first data or the second data in the document to allow registration of the position on the hierarchical structure, upon detecting that the displayed first data or the displayed second data is selected.

2. The computer-readable recording medium according to claim 1, wherein the first data and the second data that are separated by the predetermined tag or symbol are data corresponding to a title in a table as well as data corresponding to a cell relating to the title, the first data and the second data being interposed between tags indicating a start and an end of the table.

3. The computer-readable recording medium according to claim 1, wherein the first data and the second data that are separated by the predetermined tag or symbol are data corresponding to a word in a list as well as data corresponding to an explanation relating to the word, the first data and the second data being interposed between tags indicating a start and an end of the list.

4. The computer-readable recording medium according to claim 1, wherein the predetermined tag is </DT><DD> or </TH><TD>.

5. The computer-readable recording medium according to claim 1, wherein the predetermined symbol is / or :.

6. A data acquisition method that causes a computer to execute a process comprising:

extracting first data and second data that are separated from each other by a predetermined tag or symbol, from a document written in HTML;

displaying the extracted first data and the extracted second data in an associated manner; and

specifying a position of a tag included in the document on a hierarchical structure with respect to the first data or the second data in the document to allow registration of the position on the hierarchical structure, upon detecting that the displayed first data or the displayed second data is selected.

7. The data acquisition method according to claim 6, wherein the first data and the second data that are separated by the predetermined tag or symbol are data corresponding to a title in a table as well as data corresponding to a cell relating to the title, the first data and the second data being interposed between tags indicating a start and an end of the table.

8. The data acquisition method according to claim 6, wherein the first data and the second data that are separated by the predetermined tag or symbol are data corresponding to a word in a list as well as data corresponding to an explanation relating to the word, the first data and the second data being interposed between tags indicating a start and an end of the list.