Automatic Construction of a Collection of Web Page

Kenji Hirano (210076)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 13, 2004

Keywords: Web surfing support, Text classification, Collection of web link.

In this paper, the user who performs web search may acquire information simply needed from the web,a user demands it which collected the web pages that relevant to the keyword, and it aims at automatically generating a collection of links. Moreover, only by showing the web page automatically collected, a user have to visit web pages one by one in order to know where useful information is. Then, the description, about web pages are also shown. In this paper, the description about a page (henceforth, forget page) is not extracted from the web page itself to carry to a link collection, but it is extracted from the page (henceforth, reference page) which has crawled on the link at the web page to carry. Thereby, the third person's description which dose not exist in the web page itself can be shown to a user. Furthermore, the description about a web page is classified into one of several categories suchas explanation, evaluation, etc., and they are shown intelligibly for a user. Here, although collecting and sorting web pages are also important when a collection of link is shown to a user, this paper, aims at providing on descriptions of explanation and evaluation of a link page.

The overview of the system is shown below.

1. Keyword input by the user
2. Collection of URL of the page related with the keyword

3. Collection of reference page of collected target pages

4. Extraction of the description about the target page from reference pages

5. Classification to explanatory note, evaluation sentence, etc. of obtained description

6. Output of collection of links

At first,the step 1-3 of are conducted. Specifically, a user inputs a keyword. A keyword is passed to "Goo" and URL of the target page related eith the keyword is collected.Next, the reference pages of each object page are collected.The methods of collection are collected using link place URL reference of Goo.

Next, the step 4 is described. In this paper, the information about a web page is extracted by two methods using HTML tags and the site name.

The method using HTML tags extracts the information on the web page around an anchor which is in a key about a list tag, br tag, a table tag, etc. In the case of a list tag, from "li" tag in front of an applicable anchor to "li" tag behind an applicable anchor, as information about a web page is extracted .The "br" tag extracts character sequence as information about a web page, when"anchor + character sequence + br" is located in a line 3 times or more. If character sequences other than an anchor are described by the cell on the right-hand side of the cell containing an anchor in the case of the table tag, the character sequence will be extracted as a reference part. Moreover, when an anchor and a character sequence are in the same sequence of a table by turns, the character sequence in the cell under an applicable anchor is extracted as information about a web page. The method which made the site name, at first, the key specifies the site name of a target page. The character sequence in the anchor of a forget page is extracted as a site name. However, a long character sequence is not extracted as a site name. Furthrmore, since the character sequence which shows a site name is not always unique, multiple site names are extracted from two or more anchors. Next, the information about a web page is extracted using a site name Information about a web page is extracted

from the HTML tag which is behind the character sequence of a site name to the HTML tag the front character sequence of a site name.

The step 5 is the followings. In this paper, the description about the forget page extracted from the reference page is classified to one of the following categories:“evaluation: convenience” ,“evaluation: amount-of-information” , “evaluation: each” ,“explanation: function” and “explanation: description”. For each pattern , clue words are prepared. The category “ evaluation: convenience” stands for descriptions about convenience and user-friendliness of a web page, “evaluation : amount-of-information” stands for descriptions about the scale and the amount of information of a web page, “ evaluation: others ” stands for descriptions including the evaluation of those other than the amount of information and convenience, “explanation : function” stands for description about the function of a web page, and “explanation: description” stands for the descriptions about explanation of the page written except the function of a page. The automatic classification of a category was done by pattern matcing except for “explanation: description”. Furthermore, when a pattern match for four categories failed, the category of the description was considered as “explanation: description”.

In the system of this paper was evaluated experimentaly. The automatic classification of a category was conducted to 467 description extracted The precision was 0.8519 and the recall was 0.7667 for the category“evalution: convinience”. The precision was 0.8824 and the recall was 0.7692 for the category“evaluation: amount-of-information”. The precision was 0.3125 and the recall was 0.4167 for the category“evaluation: each”. The precision was 0.6442 and the recall was 0.7614 for the category“explanation: function”. The precision was 0.7797 and the recall was 0.7302 for the category“explanation: description”. Finally, the overall precision and recall was 0.7410 and 0.6617, respectively.