Automatic Generation of Link Collection on World Wide Web

Masaki Tamura (410080)

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

February 9, 2003

Keywords: WWW, Portal site, Link collection, Ambiguity of keyword, Clustering.

By the spread of World Wide Web (WWW) recently, we can get various information on the web and everyone can set up a web site easily. Therefore, huge information has been accumulated on the web. However, finding useful pages is difficult because various information exists in confusion on WWW. In this context, the portal site supports the access to WWW. However, portal sites that a user wants do not always exist in WWW. Automatic generation of the portal site that a user interests is expected.

In this paper, we try to automatically generate a link collection as one of the contents of the portal site. A link collection is a collection of link to web pages relate to the theme given by a user.

When generating a link collection, we pay attention to the ambiguity of keywords. In case that a keyword has several meanings, the correct meaning cannot be judged. For example, when "MATSUI" is given as a keyword, the system cannot understand whether correct meaning of the keyword is "MATSUI Hideki", "MATSUI Kazuo" or others. So, we propose the method that the system judges the ambiguity of meaning of a keyword automatically and create link collections for each meaning.

The processing is as follows: (1) accepting the theme, (2) getting candidate pages, (3) adding candidate pages, (4) removing unnecessary pages,

(5) clustering, (6) output. At the step (1), a user inputs one or more keywords as the theme of the link collection. These keywords are nouns. At the step (2), the system obtains candidate pages that contain keywords. The system uses Goo search engine and obtains 500 pages from the search results as candidates of the link collection. At the step (3), some pages that cannot be obtained in step (2) though they relate to the theme are added by following links in link collections. We find link collections by pattern matching. At the step (4), some pages consisting only of link collections are removed because following a link collection to other link collection put a user much efforts. At the step (5), the system performs clustering that differs from past methods of clustering pages belong to the same topic and pays attention to ambiguity of keywords. First, the system clusters some pages having same nouns that appear in before and after a keyword. Because we consider the meaning of a keyword is defined by nouns that appear before and after this keyword. Then, "base clusters" are decided by high ranked number of pages included in each cluster. Secondly, the system calculates a similarity between base clusters and pages which don't belong to base clusters. If the similarity between the cluster C and the page P is more than a threshold, the system includes P in C. We use the cosine measure as this similarity and 50 surrounding words of keywords as the feature of a page. Also, we use "TF-ICF", the product of "term frequency" (TF) and "inverse cluster frequency" (ICF) as term weights of features. ICF is the weight that characterizes each cluster. Finally, at the step (6), the system outputs a link collection based each cluster created at step (5).

According to our experiment, the precision was 67.7% and the recall was 55.1% about the detection of link collections by the step (3) and (4). There were some base clusters we can refer the meaning of a keyword well like "MATSUI Hideki", "MATSUI Kazuo" (when the keyword is "MATSUI"), "Professional Baseball" and "High-School Baseball" (when the keyword is "Baseball"), but some are not. The precision of base clusters, the ratio of correct pages that should be included in the link collection created by our method in base cluster, is 42.6%. When we evaluated only the initial candidate pages obtained at the step (2), the precision is 49.4%. On the other hand, in case that we evaluated only the added candidate pages at

the step (3), the precision is 34.5%. Moreover, the precision of clustering, the ratio of number of web pages where a keyword has same meaning as one in pages in base cluster, is 49.6%. When we evaluated only the initial candidate pages obtained at the step (2), the precision is 48.7%. On the other hand, in case that we evaluated only the added candidate pages at the step (3), the precision is 50.6%.