Automatic Generation of a Web Directory for Regional Information

Yousuke Ohtsuki

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

February 15, 2000

Keywords: Web Directory, Automated Editing, Regional Information, World Wide Web, Information Extraction.

As World Wide Web (WWW) becomes the worldwide information medium, it is now indispensable for both official organizations and individuals to announce the information. Inevitably, there is already a lot of useful regional information in WWW—from information about public service for the specific regions to information for tourists. One can travel everywhere without wondering where to go and what to see if he can use WWW as travel information source instead of printed travel magazine or tourist books. Regional information on the WWW is expected to be used as education tools at elementary schools.

There are two types of tools that help us find that regional information: search engine and web directory. Both of them have problems. A search engine returns too many pages as a search result. A web directory requires a lot of efforts to create and maintain.

As a solution to the problems, I propose a system that automatically generates a web directory for regional information.

The generated directory presents regional information for all 3427 regions (prefectures, cities, towns, and villages) in Japan. The directory provides two views: regional view and category view. In the regional view, the directory provides a page for each region. For example, the directory provides the page that contains regional information of Ishikawa prefecture. In the category view, the directory generates a page that presents information of the specific category in one or more region according to the request. For example, we can see the page about hot springs in Hokuriku area.

The system consists of the following three elements:

1. **contents database**
   The contents database stores all of the information that is presented by the directory.

2. **presentation module**
   The presentation module accepts a user's request and generates the page by retrieving the contents database.

3. **automated editing module**
   The automated editing module is the heart of the system. This module collects and generates all of the information that is presented by the directory.

The automated editing module produces the contents of the directory from these two types of information sources:

1. web site for a specific region
   This type of site presents various information of a specific region. A typical example of this type is an official web site of a city, town, and village.

2. web site for a specific type of information
   This type of site presents a specific type of information for all regions in Japan. A typical example of this type is the web site that provides population of each region.

The editing process differs by the above two types of information sources. For the first type, the editing process consists of automatic collection of web sites and automatic classification of web pages. For the second type, the editing process consists of information extraction and information reproduction.

Because there are many web sites that present information for the specific regions, it is almost impossible to collect these web sites by hand. The automatic collection of web sites enables to automate this process. The automatic collector first generates a set of URLs by a rule—the naming rule of the domains for cities and towns in Japan. By using a search engine, the collector searches the pages that contain many links to these URLs: the pages may be link collections of regional web sites. The collector extracts every unknown URL on the pages, and adds it to the set of URLs. By iterating this process, the collector collects many URLs of regional web sites. In the experiment, the collector collects 4012 URLs of regional sites in total; they cover 2852 regions (83.2 percent) of all regions in Japan.

The automatic classification assigns one of eight categories to each web page in each regional site. The eight categories are:

- general
- industry and development
- festivals and other events
- culture, history and education
- leisure and sightseeing
- statistics

- residents only
- links

The automatic classifier uses the words and phrases that are distinctive to each category. If one of these words and phrases appears in the anchor text, the page title, or other enhanced strings in the page, the classifier assigns the category associated with the words and phrases. In the closed test, we obtained 87.9 percent recall and 87.9 percent accuracy. In the open test, we obtained 71.4 percent recall and 76.2 percent accuracy.

From the second type of information source, i.e., web site for a specific type of information, the automated editor extracts information in the form of attribute-value pairs. Usually, these sites use the table format to present information of all regions: the table analysis enables extracting information from the tables. The numerical values are used for information reproduction: for example, population density is calculated from population and area.