Automatic Generation of Address Lists from the World Wide Web

Tomoki Tsuda

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

February 15, 2000

Keywords: Address Information, Address List, World Wide Web, Information Complement, Information Extraction.

The World Wide Web (WWW) is already a popular medium to announce and publish information. There are many official and unofficial web pages that describe about various things. As the number of these pages grows rapidly these years, a lot of address information (postal mail address and phone number) becomes available. In coming a few years, more and more address information will be on web pages.

There are two types of pages that contain address information:

1. Home page
   A home page of an organization such as university, library, hotel, and store, is designed to introduce itself. The page contains the address information: postal mail address and phone number.

2. Address list page
   WWW includes many small directories for the specific categories, such as university, library and hotel. Each directory provides an address list for a specific category.

Usually, we use a search engine to find the exact page that contains the address information that we want. However, a lot of pages will be listed as the search result. We have to scan these pages to obtain the address information. This is the common problem that we encounter when we use WWW.

In order to solve the problem, I propose a system that generate address lists from WWW automatically. This system accepts a category and a region as an input and generates the address list of the category in the region. For example, the system generate the address list of libraries (category) in Ishikawa prefecture (region).

The system has the following characteristics:

- The system collects address information automatically.
- The system use address list pages as the information source.
- The system extract a set of address information (zip code, postal mail address, phone number) for each organization and its category such as library or university.
- The system complement missing parts of address information automatically.

The system consists of three modules:

1. Collection module
2. Address information database
3. Retrieval module

The collection module, which is the heart of the system, consists of two submodules: page collection module and information extraction module.

The page collection module accepts a category name as an input and collect web pages that contain address list of the category by using search engines. In order to collect the pages efficiently, the module generates a set of search queries, which contain not only the input category name but also the word such as "list" and "directory". The module downloads top 100 pages for each query.

The information extraction module analyzes each downloaded page and extracts address list on the page. The module first extract tables (described by the table tag) and its headings on the page. Then, the module standardize the table format because there are various table formats to describe address lists. Finally, the module extracts each record of the table and store it to the address information database. In some cases, a few field of a record are missing. For example, in postal mail address, the city name is omitted; in phone number, the area code is omitted. The module complement these missing information by using other field of the table and the table heading.

The address information database stores all of extracted information. A record of the database consists of the following fields:

1. Name
2. Category
3. Zip
4. Postal mail address
5. Phone number
6. Extraction page title
7. Extraction page URL

8. Official page URL

The retrieval module retrieve the database according to the user request and generate the address list from the retrieval results.

I tested the system for 30 categories:

> University, High school, Junior high school, Elementary school, Kindergarten, Nursery school, Library, Art gallery, Museum, Hospital, Drugstore, Shrine, Temple, Hotel, Guest house, Inn, Restaurant, Eating house, Noodle restaurant, Bar, Skiing ground, Golf course, Sports shop, Fishing shop, Bookstore, Computer shop, Boutique, Car dealer, Garage, Hairdresser

In total, the system collected 32000 records. The 90 percent of the records have the correct category and The 79 percent of the records are correct.