A Person Finder on the World Wide Web

Ayumi Yamamoto

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

February 15, 2000

Keywords: World Wide Web, automatic extraction of people's information, table analysis, information extraction, search engine

The World Wide Web is enormous information source that contains a wide range of information. Search engines are widely used for information hunt. A search engine provides keyword search to a large number of web pages that are collected by a software called robot or spider.

These search engines help us find the web pages that contain the desired information. Because these search engines are designed for general purposes, we often find too many web pages listed as the search result, and most of these pages are not appropriate for our needs — some pages lacks the necessary information, and other pages include information that is totally nonsense to us. As a result, we can not obtain the exact information that we want.

A solution to this problem is a search system for a specific category: Limiting the category of the search targets gives two benefits:

1. It provides the filtering method of unnecessary information.

2. It determines the kind of information should be provided to users.

According to the above idea, I have implemented a person finder system which accepts a person name as an input and responses who she is. By using this system, we can easily obtain the person information that we want to know.

The system consists of four modules:

1. user interface
2. person-information database
3. off-line collection module
4. on-line collection module

The last two modules are the heart of the system.

The off-line collection module accepts an occupation category such as Seijika (politicians) and Chojitsuka (writers), and collects the people's list which contains personal information of that occupation category. The collection process consists of two steps: page collection and table analysis.

The first step collects the pages that contain the people list of the given occupation category by using search engines and anchors (hyperlinks). First, this step sends a set of search queries that contain the given category and the word such as "list", and downloads the top 300 pages for each query. Second, the step analyses each downloaded page and extracts all anchors on the page. When an anchor text indicates the pointed page may contain the people's list, the step downloads the page. Finally the step checks whether the people's list exists or not for each downloaded page. If it exists, the page becomes the candidate for the next step.

The second step extracts people information from each candidate page. First, the step extracts the people's list (in table format) and its heading. Second, the step standardizes the table format, because there are various table formats to describe people's lists, Third, the step extracts personal properties such as name and birthday for each person in the list. Finally, the step extracts the occupation subcategory from the table heading. All of extracted information are stored in the person-information database.

The on-line collection module accepts the person name as an input and generates the profile (in text form) of that person as the output. The collection process consists of two steps: page collection and layout analysis.

The first step collects the pages that contain the personal profile of the given person name by using search engines. This step sends a set of search queries that contain the given person name and the word such as "birth", and downloads the top ten pages for each query.

The second step analyses each downloaded page and extracts the profile. First, the step simplifies the page layout and markups the layout properties such as link break and indent. Second, if the step finds the heading that contains the given name, it extracts the text associated with the

heading as the profile.

I conducted two experiments on person information collection: an experiment with off-line collection module and another experiment with on-line collection module.

To evaluate the off-line module, I used two occupation categories: Seijika (politicians) and Chojitsuka (writers). In case of Seijika, the module collects personal information of Kokkai-Giin (members of Diet) or Chihou-Giin (prefecture councilors, city councilors, etc.). In case of Chojitsuka, the module collects personal information of award-winning writers. The following results were obtained:

1. The module collected personal information for 2245 persons in case of Seijika. However, it collected information for only 833 persons in case of Chojitsuka.

2. The module collected not only general properties such as name and birthday but also occupation-specific properties, e.g., Senkyo-ku (constituency) of Seijika.

To evaluate the on-line collection module, I used fifty names of Sangiin-Giin (members of the House of Councilors) and 128 names of Chojitsuka. I obtained the following results:

1. The module succeeded to find personal profiles for 104 people out of 128 people of Chojitsuka. However, it succeeded to find profiles for fourteen people out of fifty people of Sangiin-Giin.

2. 79 percent of the extracted profiles were appropriate.