Computational Linguistics and Chinese Language Processing
vol. 3, no. 1, February 1998, pp. 75-100

White Page Construction from Web Pages for Finding People on the Internet

Hsin-Hsi Chen* and Guo-Wei Bian*

Abstract

This paper proposes a method to extract proper names and their associated information from web pages for Internet/Intranet users automatically. The information extracted from World Wide Web documents includes proper nouns, E-mail addresses and home page URLs. Natural language processing techniques are employed to identify and classify proper nouns, which are usually unknown words. The information (i.e., home pages' URLs or e-mail addresses) for those proper nouns appearing in the anchor parts can be easily extracted using the associated anchor tags. For those proper nouns in the non-anchor part of a web page, different kinds of clues, such as the spelling method, adjacency principle and HTML tags, are used to relate proper nouns to their corresponding E-mail addresses and/or URLs. Based on the semantics of content and HTML tags, the extracted information is more accurate than the results obtained using traditional search engines. The results can be used to construct white pages for Internet/Intranet users or to build databases for finding people and organizations on the Internet. Such searching services are very useful for human communication and dissemination of information.

Keywords: proper name identification, information extraction, white pages, World Wide Web

1. Introduction

With the rapid growth of the Internet in recent years, the World Wide Web (WWW) has become a powerful medium for human communication and dissemination of information. Because more online information is disseminated through this giant media, the Web forms a very large knowledge resource. The explosive growth of the WWW has involved more than 10 million documents. Some search engines and information discovery systems have been introduced to help users locate relevant information. However, one

*Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. E-mail: hh_chen@csie.ntu.edu.tw

of the problems in cyberspace is that it is very difficult to know how to contact an entity, which is a concrete object that can send and receive information. For communication purposes, we usually want to know a person's or a company's E-mail address, or his/her home page URL. White pages, which are E-mail directories or URL directories in this case, can help users find such information. There are two major difficulties in building Internet white pages and people searching services. First, it is hard to set up such a white page manually because the WWW is a very large database and is created dynamically. Second, the approaches adopted by current search engines are not suitable for searching the e-mail addresses and home page URLs of people and organizations.

Current search engines only index the contents (words) of a web page with the page's URL. When a page contains many proper names, the search engine will index all the proper nouns with this page's URL. However, only one of these proper nouns or none is the owner of the page. Example 1(a) shows the appearance of a portion of a web page (http://www.ntu.edu.tw/TANet/public.html). Totally, there are 42 public universities and colleges listed in this page. The traditional search engines will index all of these 42 proper nouns with the page's URL, but none of the home pages of these proper nouns is this page.

Example 1(a). http://www.ntu.edu.tw/TANet/public.html (Browsing View)

公立大學暨獨立學院
Public University and College
國立臺灣大學 / National Taiwan University
台北市(10764)羅斯福路四段 1 號
1, Roosevelt Rd., Sec.4, Taipei, Taiwan, R.O.C.
Tel.:(02)3630231
Fax :(02)3627651
國立政治大學 / National Chengchi University
台北市(11623)指南路二段 64 號
64, Chih-Nan Rd., Sec.2, Taipei, Taiwan, R.O.C.
Tel.:(02)9393091
Fax :(02)9398043
國立清華大學 / National Tsing Hua University
新竹市(30043)光復路二段 101 號
101, Kuang-Fu Rd., Sec.2, Hsinchu, Taiwan, R.O.C.
Tel.:(035)715130
Fax :(035)722467
國立交通大學 / National Chiao Tung University
新竹市(30050)大學路 1001 號
1001, Ta-Hsueh Rd., Hsinchu, Taiwan, R.O.C.
Tel.:(035)712121
Fax :(035)721500
……...

Example 1(b). http://www.ntu.edu.tw/TANet/public.html (Original HTML)

```
<html>
<head>
<title>公立大學暨獨立學院 / Public University and College</title>
<!--版權所有：國立臺灣大學對此文件保留所有權利-->
<!--Copyright (c) 1996 National Taiwan University ALL RIGHTS RESERVED-->
</head>  <p>
<h2>公立大學暨獨立學院<br>Public University and College</h2>
<ul type=square>

<li><a href="http://www.ntu.edu.tw/">國立臺灣大學 / National Taiwan University</a>
   <ul type=disc>
        <li>台北市(10764)羅斯福路四段 1 號
        <li>1, Roosevelt Rd., Sec.4, Taipei, Taiwan, R.O.C.
        <li>Tel.:(02)3630231
        <li>Fax :(02)3627651
   </ul>
<li><a href="http://www.nccu.edu.tw/">國立政治大學 / National Chengchi University</a>
   <ul type=disc>
        <li>台北市(11623)指南路二段 64 號
        <li>64, Chih-Nan Rd., Sec.2, Taipei, Taiwan, R.O.C.
        <li>Tel.:(02)9393091
        <li>Fax :(02)9398043
   </ul>
<li><a href="http://www.nthu.edu.tw/">國立清華大學 / National Tsing Hua University</a>
   <ul type=disc>
        <li>新竹市(30043)光復路二段 101 號
        <li>101, Kuang-Fu Rd., Sec.2, Hsinchu, Taiwan, R.O.C.
        <li>Tel.:(035)715130
        <li>Fax :(035)722467
   </ul>
<li><a href="http://www.nctu.edu.tw/">國立交通大學/ National Chiao Tung University</a>
   <ul type=disc>
        <li>新竹市(30050)大學路 1001 號
        <li>1001, Ta-Hsueh Rd., Hsinchu, Taiwan, R.O.C.
        <li>Tel.:(035)712121
        <li>Fax :(035)721500
   </ul>
   ……..
</html>
```

Nevertheless, the original HTML data shown in Example-1(b) shows some information about these proper nouns. Some anchor tags are provided for users to browse their home pages. For example, the URL of the home page of National Taiwan University (國立臺灣大學) is http://www.ntu.edu.tw/, which is described by the HTML anchor tag. Considering the context and HTML tags, these proper nouns and the associated anchor tags can be extracted from the web page. For example, 'National Tsing Hua University' (' 國立清華大學 ') should be related to the URL http://www.nthu.edu.tw/. On the contrary, the traditional search engines will index this school and others with the web page's URL (http://www.ntu.edu.tw/TANet/public.html). In terms of accuracy, the approach adopted by the current search engines is not suitable for such a task of finding

people and organizations. Additionally, most of the anchors and the contents in a web page are not proper nouns. Such anchors and words should not be extracted and indexed for people searching.

Furthermore, some proper nouns may appear in the non-anchor part of a web page. If there are URLs of web pages and e-mail addresses on the same page, the relationships between the proper nouns and the information should be identified. Example 2 shows such a case. In contrast to the anchor part, no explicit HTML tags indicate the relationship between the proper nouns and e-mail addresses.

Example 2 - 區域網路線路維護流程

北區區域網路中心(臺大計中網路組)

……
網路組成員:

組長: 游張松 教授
E-mail: yucs@ccms.ntu.edu.tw
Tel: 3627734 ext 219

組員: 胡 湘 小姐	李光偉 先生
E-mail: giraffe@ccms.ntu.edu.tw	edward@ccms.ntu.edu.tw
Tel: 3627734 ext 241	3627734 ext 241
組員: 曾珀雯 小姐	徐信權 先生
E-mail: popo@ccms.ntu.edu.tw	kevins@ccms.ntu.edu.tw
Tel: 3627734 ext 241	3627734 ext 241

……

How to identify the proper nouns in a web page is a critical problem for building white pages. Fortunately, a very large portion of the WWW is composed of natural language documents, which can be regarded as a text corpus. Corpus analysis techniques in natural language processing [CL, 1993] can be employed to extract knowledge from the WWW. And using the semantics of the content and HTML tags, the information (URLs and e-mail addresses) can be related to proper nouns. This paper will propose a method to construct white pages for Internet/Intranet users automatically. It extracts information, including proper nouns, E-mail addresses and home page URLs, from WWW documents, and finds the relationships among these data. The problems to be tackled are as follows:

(1) Proper nouns, which are always unknown words, have to be identified and classified from a WWW corpus. Personal names and organization names are the requested entities for people finding on the Internet. Those proper nouns that denote organizations are usually hierarchical. Such relationships must be distinguished.
(2) There may be more than one proper noun, more than one E-mail address, and more

than one URL in a WWW document. Thus, we have to find a mapping from a set of E-mail addresses (or URLs) to a set of proper nouns.

The extraction method proposed in this paper was tested on the web pages in Taiwan. Section 2 introduces WWW documents and the semantics of the HTML annotation. The hierarchical nature and the related HTML tagging (1996) are discussed. Section 3 gives an overview of our white page constructor. Section 4 presents the identification algorithms for proper nouns. Here, we focus on personal names and organization names. Section 5 touches on the algorithms for mapping between proper nouns and related information. Section 6 discusses the experiments, and Section 7 offers some conclusions.

2. WWW Documents

The first step in constructing white pages is to find out where proper nouns, E-mail addresses and URLs are located in WWW documents. Web documents are different from a traditional text corpus in that they are HTML (HyperText Markup Language) files. The tagging information provides some clues, but it also introduces some noise. How to use the information is a very important issue in applications on the Internet, e.g., cross-language information retrieval [Bian and Chen, 1997]. In plain text, each sentence always has a sentence terminator, such as a full stop, question mark or exclamation mark. These symbols split each document into several processing units. In HTML files, these punctuation marks do not always appear. Quasi-sentences are defined according to some HTML tags shown below:

- Title (TITLE)
- Headings (H1, H2, ⋯ , H6)
- Address (ADDRESS)
- Unordered Lists (UL, LI)
- Ordered Lists (OL, LI)
- Definition Lists (DL, DT, DD)
- Tables (TABLE, TD, TH, TR)

Furthermore, some punctuation symbols like '|' and ':' have the same effects. In contrast to the above sentence delimiters, the font style elements may introduce noise. Bold (B), italic (I), superscripts (SUP), subscripts (SUB) and font (FONT) can be used to emphasize some points in texts. However, these elements produce many unknown words because a word is split into several parts by HTML tags. Example 3 illustrates the word 'Font' associated with various font style tags. Thus, these tags should be treated as meta-information and hidden from processing.

Example 3.

```
<I><B><FONT SIZE=+2><FONT COLOR="#FF0000">F</FONT>
<FONT COLOR="#0000FF">o</FONT>
<FONT COLOR="#FF8000">n</FONT>t</FONT></B></I>
```

Links denoted by anchors (A) in WWW documents are possible sources of proper nouns and related information. The WWW documents shown in Appendix A shows their typical features. The first example is the home page of National Taiwan University (NTU, http://www.ntu.edu.tw/). The entity that we are interested in is 'National Taiwan University' (' 國立台灣大學 '), which is an organization name and is shown in the title area. The second example (http://www.ntu.edu.tw/NTULink/) follows from 'NTU Link' on the NTU home page. An underline shows a link to other home pages in the web page. The interesting entities are Office of Academic Affairs (' 教務處 '), Office of Student Affairs (' 學務處 '), and Office of Business Affairs (' 總務處 '); University Library (' 圖書館 '); Computer and Information Network Center (' 計算機及資訊網路中心 '); Population Studies Center (' 人口研究中心 '). Those units that do not have any links are not considered. For example, the home pages for Accounting Office (' 會計室 ') and Military Instructors' Office (' 軍訓室 ') have not yet been constructed now, so that they are not listed in the final white pages. Following the link for 'Colleges, Schools, Departments, Graduate Institutes and Affiliated Organizations', we can retrieve more information. All these units form a hierarchical structure in National Taiwan University.

A link in the HTML file may be represented as follows:

 text

When "text" is a proper noun, its home page URL can be described by "argument". Consider an example on the 'NTU Link' web page. The link to 'Office of the Dean of Academic Affairs' (' 教務處 ') is shown below:

 教務處
/ Office of the Dean of Academic Affairs

If the proper noun and its URL are put into white pages directly, this entry may be ambiguous. This is because many universities have similar organizations. Therefore we should keep the hierarchical path of the web page to disambiguate the meaning of a proper noun. Further, the relative URLs have to be changed into absolute ones to keep all of the URL information. Because the URL associated with the link 'Office of the Dean of Academic Affairs' (教 務 處) is a relative URL

(/Campus/announce/index.html#academic) and the web page's URL is http://www.ntu.edu.tw/NTULink/, the absolute URL of this organization is represented as http://www.ntu.edu.tw/Campus/announce/index.html#academic. In addition, the host name (www.ntu.edu.tw) in the hierarchical path of this URL shows that this organization is part of National Taiwan University. The complete organization name will be 'Office of the Dean of Academic Affairs in National Taiwan University' (' 國立台灣大學 教務處 '). Therefore, similar organizations and personal names can be disambiguated with the host names of their absolute URLs to find their home pages' URLs on the global Internet.

Besides the linking anchor field, proper nouns may appear in other portions of a WWW document. Dealing with these objects is more complex because no explicit HTML tags indicate the URLs of these objects. An additional algorithm is needed to associate URLs and E-mail addresses with suitable proper nouns. Different kinds of clues, such as the spelling method, adjacency principle and HTML tags (e.g., title, headings, address, and font style elements), are employed.

3. System Overview

We periodically collect web pages from the Internet/Intranet using a spider. The white page constructor first analyzes these HTML files. Basic processing units (sentences or quasi-sentences) and HTML meta-information are gathered. Because a Chinese sentence (or quasi sentence) is composed of a sequence of characters without word boundaries [Chen and Lee, 1996], a Chinese segmentation system identifies the word tokens. Then, a proper noun identification system (see Section 4) extracts personal names and organization names. During processing, the information in the anchor parts is placed in the anchor set (AS). Other information, i.e., that appearing in non-anchor parts, is placed in one of the content sets (CSes) which correspond to different types of information. In the current implementation, there are three content sets: CS_Proper-Noun, CS_E-Mail and CS_HTTP. They record proper nouns, E-mail addresses and URLs, respectively. For the anchor set, the remaining task is simple. We just relate the proper noun found in an anchor to the corresponding URL or E-mail address. For the content sets, a mapping algorithm (see Section 5) is proposed to associates URLs and/or E-mail addresses with a suitable proper noun. Algorithm 1 shows the information extraction part of the white page constructor.

Algorithm 1. Information Extraction

Input: An HTML file or a plain text with its URL (URL_1)

Output: An anchor set (AS) and three content sets (CSs)

Method:

1. [HTML Parser]
   Identify sentence boundary and collect those HTML tags that are useful for information mapping.

2. [Chinese Segmentation System]
   For each processing unit (a sentence or a quasi-sentence), identify the word boundary.

3. [Identification of Proper Nouns]
   Identify and classify proper nouns in the text.

4. For each proper noun (PN)
   {
   4.1 [Analyze the <Title> tag: <TITLE>Title_Text</TITLE>]
   if PN tagged with the HTML tag <Title>,
   add the tuple (PN, URL_1) to the Anchor Set (AS)
   4.2 [Extract the Anchor Information]
   if PN tagged with the HTML tag <A>
   (Text),
   add the tuple (PN, protocol://host/path) to the Anchor Set (AS)
   4.3 [Extract the Content Information]
   if PN is in the non-anchor part (content)
   add PN to CS_Proper-Noun with the following attributes:
   the position information of token (token_no) and the associated HTML meta information (<TITLE>, <Hn>, <Address>, <Bold>, <Font> and <Italic>)
   }

5. Extract different types of information with the position information of token (token_no), and add to the corresponding Content Sets (CS_E-Mail and CS_HTTP)

6. End

4. Identification of Proper Nouns

Proper nouns that are not collected in lexicons are major unknown words in natural language texts. Several methods [Boguraev and Pustejovsky, 1996; Mani, *et al*., 1993; McDonald, 1993; Paik, *et al*., 1993] have been proposed to identify English proper nouns. For research related to Chinese, Chang *et al*. [1992] and Wang *et al*. [1992] touched on Chinese personal names; Sproat *et al*. [1994] considered Chinese personal names and transliterations of foreign words; Chen and Lee [1996] identified Chinese personal names, Chinese transliterated personal names and organization names. The name identification module is based on our previous design. The methods are described below.

4.1 Identification of Personal names

A Chinese personal name is composed of surname and given name parts. Most Chinese surnames are single characters (model (a)), and some rare ones have two characters (model (b)). A married woman may place her husband's surname before her surname (model (c)). Thus there are three possible types of surnames, i.e., single character, two characters and two surnames together. Most names have two characters, and some rare ones are single characters. Theoretically, every character can be considered as a names rather than a fixed set. Thus, the length of Chinese personal names ranges from 2 to 6 characters. The baseline models for identification are shown as follows.

Model (a) Single character surname:

$$(1)\ \frac{\#C_1}{\&C_1}\times\frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold1$$

$$(2)\ \frac{\#C_1}{\&C_1} > Threshold2 \text{ and } \frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold3$$

Model (b) Two characters surname:

(3) $C_{11}C_{12}$ is two-character surname and $\frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold4$

Model (c) Two surnames together:

$$(4)\ \frac{\#C_{11}}{\&C_{11}} \times \frac{\#C_{12}}{\&C_{12}} \times \frac{\#C_{2}}{\&C_{2}} \times \frac{\#C_{3}}{\&C_{3}} > Threshold5$$

$$(5)\ \frac{\#C_{11}}{\&C_{11}} \times \frac{\#C_{12}}{\&C_{12}} > Threshold6 \text{ and } \frac{\#C_{2}}{\&C_{2}} \times \frac{\#C_{3}}{\&C_{3}} > Threshold7$$

where C_1, C_{11}, and C_{12} are the characters forming a surname,

C_2 and C_3 are the characters which are considered as names,

$\#C_i$ is the frequency of C_i being a surname or a name,

$\&C_i$ is the frequency of C_i being contained in the other words.

For different types of surnames, the different models are adopted. Because the two-character surnames are always indicated as surnames, Model (b) neglects the score of the surname part. Models (a) and (c) have two score functions. They solve the problem of very high scores of surnames. The above three models can be extended to single-character names by ignoring the last character C_3 in each formula for training and testing. When a candidate cannot pass the thresholds, its last character is cut off and the remaining string is tried again. The frequencies of characters being surnames or names are trained from a large-scale Chinese name corpus of 219,738 Chinese personal names and 661,512 characters. The frequencies of characters being other words are trained from an NTU balanced corpus to compute the variation of characters. In total, this corpus has 113,647 words and 191,173 characters. Thresholds are trained using the Chinese name corpus. We calculate the scores of all Chinese personal names in the corpus using the above formulas. The scores for each formula are sorted, and then the one that is less than 99% of the personal names is considered to be a threshold for this formula. That is, 99% of the training data can pass the threshold.

Chinese personal names are not always composed of single characters. For example, the name part '聰明' (Cong-ming) of the sentence '陳聰明 醫術 非常高明' (Chen Cong-ming yishu feichang gauming; Chen Cong-ming has find command of the medical art) is a word. How to tell that a word is a content word or a name is indispensable. Mutual information [Church and Hanks, 1990], which provides a measure of word association, is employed to differentiate between a name and a content word. We check the string that can serve as a name or a content word with its surrounding words. When they have a strong relationship, it has high probability of being a content

word rather than a name. In the example ' 陳　家世　清白，絕　不會　犯法… ' (Chen jashi qingbai jue buhui fanfa … ; Chen has a clean family background and will never violate the law …), the two words ' 家世 ' (jashi) and ' 清白 ' (qingbai) have high mutual information, so that ' 陳　家世 ' (Chen jashi) is not a personal name in this example. Three newspaper corpora (total size about 2.6 million words) are used to train the word association.

Punctuation marks play an important role in identification. Personal names usually appear at the head or the tail of a sentence. A candidate is given an extra bonus when it is found in one of these two places. Gender has a special role in Chinese personal names. A married woman may place her husband's surname before her surname. That forms the personal name of model (c). Gender information helps us to disambiguate the type of personal name.

The last clue is the paragraph information. A personal name may appear more than once in a paragraph. This phenomenon is useful during identification. We use a cache to store identified candidates and reset the cache before next paragraph is processed. Consider the examples ' 焦仁和　表示　… ' (Jiao Renhe biaoshi … ; Jiao Renhe expressed …) and ' 焦仁和　秘書長　… ' (Jiao Renhe mishuzhang ...; Jiao Renhe Chief Secretary ...). Two candidates ' 焦仁 ' (Jiao Ren) and ' 焦仁和 ' (Jiao Renhe) are proposed and stored in the cache, but the personal name is finally identified as ' 焦仁和 ' (Jiao Renhe). For details, the reader is referred to a previous paper [Chen and Lee, 1996].

4.2 Organization Names

The structure of organization names is more complex than that of personal names. Basically, a complete organization name can be divided into two parts, i.e., name and keyword. Many words can serve as names, but only some fixed words can be regarded as keywords. Thus, keywords are important clue used to extract organization names. However, there are still several difficult problems. First, a keyword is usually a common content word. It is not easy to differentiate between a keyword and a content word. This problem results in ambiguities in POS tagging and word sense. Second, a keyword may appear in an abbreviated form. Third, a keyword may be omitted completely. Fourth, some organization names are very long, so it is hard to decide on the left boundary. The following examples illustrate these problems.

(1) Ambiguity of keywords:

(1.1) Ambiguity of word senses and POS tagging:

學會 (xuehui; Association or Learn)

(1.2) Ambiguity of POS tagging (verb or noun):
調查中心 (iaocha zhongxin; center of investigation), 研究中心 (yanjiu zhongxin; center of research), 開發公司 (kaifa kongsi; company of development), 開發中心 (kaifa zhongxin; center of development), 發展協會 (fazhan xiehui; development association), 規劃小組 (guihua xiaozu; planning group), 研習社 (yanxishe; research club), 評論社 (pinglunshe; discussion club), 發明社 (famingshe; invention club), 聯誼會 (lianyihui; social gathering)

(2) Abbreviated keywords:
投顧 (' 投資顧問公司 '): tougu (touziguwengongsi); Security Investment Consulting
護專 (' 護理專科學校 '): huzhuan (hulizhuankexuexiao); college of nursing
專校 (' 專科學校 '): zhuanxiao (zhuankexuexiao); college for professional training
工專 (' 工業專科學校 ') gongzhuan (gongyezhuankexuexiao); college of technology
商專 : shangzhuan; college of commerce
藝專 : yizhuan; college of arts
實小 (' 實驗小學 '): shixiao (shiyanxiaoxue); experimental primary school

(3) Keyword omitted:
宏碁 (hongji; Acer), 友訊科技 (youxunkeji, D-Link Tech.),
友力資訊 (youlizixun; Ulead Inc.)

(4) Long organization names:
國立台灣工業技術學院 (guoli taiwan gongye jishuxueyuan; National Taiwan Institute of Technology), 國家地震工程研究中心 (guojia dizhen gongcheng yanjiu zhongxin; National Center for Research on Earthquake Engineering),
實踐設計管理學院 (shijian sheji guanli xueyuan; Shih Chien College of Design and Management), 台北市大安區萬芳社區發展協會 (taibei shi daan qu wanfang shequ fazhan xiehui; Taipei Daan District Wanfang Community Development Association)

Our previous work [Chen and Lee, 1996] only touched on the fourth problem. Keywords, which are good indicators, play a role similar to that of surnames. They show not only the possibility of an occurrence of an organization name, but also its right boundary. A prefix is a good marker for a possible left boundary, for example, ' 國立 ' (National), ' 省立 ' (Provincial), and ' 私立 ' (Private), and so on. The name part of an organization may consist of single characters or words. Parts of speech, such as transitive verbs, adjectives, numerals and classifiers, are also useful for determining the left boundary. The name part of an organization cannot cross these critical parts of speech. For example, ' 公司 ' (company) in ' 三家公司… ' (three company …) is not a keyword

due to the critical parts of speech. Because a tagger is not involved before identification, the part of speech of a word is determined wholly based on its lexical probability.

Although our previous experiment has shown that these critical parts of speech are useful in determining the left boundary of an organization name in a newspaper text, the ambiguity of parts-of-speech (as verb or noun) decreases the performance for the specific task - identification of organization names in an unrestricted domain. For example, the identification system will miss or give an incorrect left boundary for organization names containing ' 調查 ' (diaocha; investigation), ' 研究 ' (yanjiu; research), ' 開發 ' (kaifa; development), ' 發展 ' (fazhan; development), ' 規劃 ' (guihua; plan), ' 研習 ' (yanxi; study), ' 評論 ' (pinglun; critique), ' 設計 ' (sheji; design), ' 管理 ' (guanli; management), ' 發明 ' (faming; invention), and so on. To resolve this problem in proper noun extraction, a refined method is proposed to deal with such organization names. The experiments described in Section 6 will illustrate the performance of the baseline and refined methods.

5. A Mapping Algorithm

Identified proper nouns may appear in the anchor parts or the non-anchor parts of HTML files. For proper nouns in anchor parts, the anchor tags indicate their home pages' URLs or e-mail addresses explicitly. Consider the example " 國立臺灣大學 / National Taiwan University ". The URL of the home page of National Taiwan University (國 立 臺 灣 大 學) is http://www.ntu.edu.tw/ as described by the HTML anchor tag. Based on the HTML tags, the information about these proper nouns attributed by anchor tags can be extracted easily from web pages.

For proper nouns appearing in non-anchor parts, a more complicated procedure is employed. Because the relationships between the proper nouns and the corresponding information are not specified explicitly, a mapping scheme can be used to associate URLs and e-mail addresses with suitable proper nouns. The following shows an example.

Example 4 - 各系所網路管理人 (the network manager of each department) (http://www.ntu.edu.tw/NTUCC/NetManager.html)

各系所網路事務之管理負責人

DEPNAME	ROUTEMAN	ROUTETEL	ROUTEOFF	EMAIL
電機工程學系、所	王凌霄	3212-4 ext 234	電機工程學系 234 室	
材料研究所	曾德玉	3638912	工綜館 625 室	
化工系、所	吳名弘	2185	化工系機算機室	
資訊系、所	黃育銘	3625336-221	資訊新館 2F221	root@csman.csie.ntu.edu.tw
……..				
地理系、所	蔡博文	2147	地理系管二樓	tsaibw@ccms.ntu.edu.tw
地質系、所	蕭銘璽	2341 ext. 13	地質系 313 室	r2204204@sun03.gl.ntu.edu.tw
動物系、所	丘台生	2128	漁科館 401 室	tschiu@ccms.ntu.edu.tw
……				

Algorithm 2 illustrates the mapping between URLs (and/or E-mail addresses) and proper nouns. A score function that considers the spelling method, adjacency principle and HTML tags is used to determine the relationships among proper nouns and related information.

The ranking function is defined as follows:

Score(Info, PN) =

$$\left(\frac{\text{HTML\_SCORE(PN)}+1}{\text{abs(Info.token\_no - PN.token\_no)}}+\frac{\text{Title(PN)}}{\text{Total\_tokens - Info.token\_no}+1}\right)$$
$$+\text{Pinyin\_Similarity(PN, Info)} * \text{E - mail(Info)} * \text{Weight}$$

HTML_SCORE(PN) =

$$\mathit{Title}(PN)+\mathit{Heading}(PN)+\mathit{Address}(PN)+\mathit{Bold}(PN)+\mathit{Font}(PN)+\mathit{Italic}(PN)$$

where Info is a URL or an e-mail address,

PN is a proper noun,

Info.token_no and PN.token_no are the positions of the specific tokens,

Total_tokens is the total number of tokens in the file,

Title(), Heading(), Address(), Bold(), Font(), Italic() and E-mail() are the Boolean functions,

Pinyin_Similarity(PN, Info) is defined below and used to measure the similarity between PN and Info under the criteria of Pinyin,

Weight is used to measure the importance of Pinyin similarity.

Algorithm 2. Information Mapping

Input: Three Content Sets (CSs)
A Threshold and a Window_Size of context

Output: A Mapping Set (MS)

Function: Mapping CS_E-mail (CS_HTTP) to CS_Proper-Name

Method:
1. Set MS to be an empty set.
2. For each CS information set (i.e., CS_E-mail and CS_HTTP)
   { /* the mapping between CS and CS_Proper-Noun may be *Many-to-one*. */
   copy CS_Proper-Noun to CD
   for each entry Info in CS
   { PN is an entry whose offset from Info is less than Window_Size, and *Score*(Info, PN) is the maximum in CD. If many entries have the same maximum value, the entry appearing before Info is chosen.
   if *Score*(Info, PN) > Threshold
   { add (Info, PN) into MS
   }
   }
   }
3. End

The *Score* function combines the following heuristic rules:

(1) **Spelling Method**. If the extracted information (Info) is an E-mail address, the similarity between Info and the proper noun (PN) is considered. Because the user-id in an E-mail address is often transliterated from a Chinese name, this heuristic rule is preferred over other cues, and we assign it a larger weight. The Pinyin system [Lu, 1995] is adopted to transliterate Chinese names. For robustness, the Pinyin similarity is defined as follows:

Pinyin_Similarity(PN, E-mail) =

$$\frac{\#\text{ of letters in user - id that match the pinyin transliteration of PN}}{\text{total \# of letters in the user - id of the E - mail address}}$$

where PN is a proper noun, and E-mail is an e-mail address.

For example, the Pinyin transliteration of " 邊國維 " is "Bian Guo Wei". The similarities between the following e-mail addresses and this personal name are:

Pinyin_Similarity(邊國維 , gwbian@nlg.csie.ntu.edu.tw) = $\frac{6}{6}$ = 1

Pinyin_Similarity(邊國維 , arthur_bian96@nlg.csie.ntu.edu.tw) = $\frac{4}{10}$ = 0.4

Pinyin_Similarity(邊國維 , arthur@nlg.csie.ntu.edu.tw) = $\frac{0}{6}$ = 0

(2) **Adjacency Principle.** Proper nouns and the related information are often close to each other. The distance between Info and PN is measured in terms of the number of intervening tokens. Recall that we assign each object a unique token number. Closer pairs have larger scores. Additionally, a proper name appearing in the title of a web page (tagged with <Title>) will be treated close to the rear of a web page.

(3) **HTML Tags.** Proper nouns (PNs) that appear in Title (<Title>), Heading (<Hn> … </Hn>) or Address, or are described by the font style (Bold, Italic and Font tag elements) are given larger weights than other normal proper nouns.

6. Experiments

In our initial experiments, a total of 703 web pages were collected from the NTU Web (http://www.ntu.edu.tw/). A person identified the personal names and organization names in these web pages and associated them with the URLs and the e-mail addresses if possible. Then, the collected answers were classified into an anchor set and a content (non-anchor) set.

The results of identification using the proposed system were checked against human results. The window size (Window_Size) of context was 6, and the score threshold (Threshold) was 0.2 for the mapping algorithm. The threshold was greater than the inverse of the window size. It was used to filter out proper nouns that were near the window boundary but were not described by any HTML tags.

Table 1 shows the results of identification in both sets and the mapping result in the content set. In Table 1(a) and 1(b), the number of personal names and the number of organization names identified by humans are listed in column 2. Columns 3 and 4 show the identification results of proper nouns using our system and the correct results. The precision and the recall are defined as follows.

$$\text{Precision} = \frac{\#\text{ of items identified correctly by program}}{\#\text{ of items identified by program}}$$

Table 1. *The Results of Identification and Information Mapping*

(a) Identification of Proper Nouns in the Anchor Set

Anchor Set	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Personal name	255	228	189	82.89%	74.12%
Organization Name	746	611	213	34.86%	28.55%

(b) Identification of Proper Nouns in the Content Set

Content Set	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Personal name	1732	3343	1470	43.97%	84.87%
Organization Name	3029	2272	503	22.14%	16.61%

(c) Identification of Organization Names Using Refined Method

Organization Name	# of items in the web pages of NTU	# of items identified by program	# of items identified correctly by program	Precision	Recall
Anchor Set	746	856	558	65.19%	74.80%
Content Set	3029	3392	2082	61.38%	68.74%

(d) The Mapping Result in the Content Set

Content Set Mapping	# of items extracted by program	# of items mapped correctly by program	# of items mapped incorrectly by program	Accuracy
E-mail	64	18	5	78.26%
HTTP	16	1	0	100%

$$\text{Recall} = \frac{\#\text{ of items identified correctly by program}}{\#\text{ of items in the web pages}}$$

In the anchor part, there were 6,204 linking items. Of these, the numbers of personal names and organization names were 255 and 746, respectively. That is, 83.87% of the anchors were irrelevant and should be screened out for the task of finding people. The precision and the recall rates were 82.89% and 74.12% for the identification of personal names, respectively.

However, the precision and the recall rates for the identification of organization names were much lower than those obtained in our previous work. The major errors resulted from the strategy discussed in Section 4.2, i.e., "parts of speech such as transitive verbs, adjectives, numerals and classifiers are also useful to determine the left boundary,

and the name part of an organization cannot cross these critical parts of speech." Many of the organization names may contain ' 調查 ' (diaocha; investigation), ' 研究 ' (yanjiu; research), ' 開發 ' (kaifa; development), ' 發展 ' (fazhan; development), ' 規劃 ' (guihua; plan), ' 研習 ' (yanxi; study), ' 評論 ' (pinglun; critique), ' 設計 ' (sheji; design), ' 管理 ' (guanli; management), ' 發明 ' (faming; invention), and so on. All of these words can be nouns or transitive verbs. The identification system misses or gives the incorrect left boundary for such an organization name. The following examples illustrate this problem.

公共政策研討學會 (gonggong zhengce yantao xuehui; Public Policy Workshop Association), 科學管理學會 (kexue guanli xuehui; Science Management Association), 中央研究院調查研究工作室 (zhongyang yanjiuyuan diaocha yanjiu kongzuoshi; Office of Survey Research at Academia Sinica), 電影研究社 (dianyin yanjiushe; Movie Club), 機車研習社 (jiche yanxishe; Motorcycle Club), 女青年聯誼會 (nuqingnian lianyihui; Youth Women's Christian Association), 舞台設計工作室 (wutaisheji gongzuoshi; Studio of Stage Design), 台北市大安區萬芳社區發展協會 (taibei shi daan shequ fazhan xiehui; Taipei Daan District Wanfang Community Development Association), 實踐設計管理學院 (shijian sheji guanli xueyuan; Shih Chien College of Design and Management), 台北影視開發公司 (taibei yingshi kaifa gongsi; Taipei Movie, Video and Television Development Company), 臺灣大學推廣教育中心 (taiwan daxue tuiguang jiaoyu zhongxin; Center of Extended Education, National Taiwan University), 水產試驗所 (shuichan shiyansuo; Fishery Research Institute), 台大校園規劃小組 (taida xiaoyuan guihua xiaozu; Campus Planning Group, National Taiwan University), 領導公關公司 (lingdao gongguan gongsi; Lingdao Public Relation Company), 編輯委員會 (bianji weiyuanhui; Editing Committee), 調查委員會 (diaocha weiyuanhui; Investigation Committee), 污泥處置研究所 (wunichuzhi yanjiusuo; Mud Disposal Research Institute), 交大應用藝術研究所 (jiaoda yingyong yishu yanjiusuo; Institute of Applied Arts, National Chiao Tung University), 生物技術開發中心 (shengwu jishu kaifa zhongxin; Development Center for Biotechnology), 中華民國視訊發展協會 (zhonghua shixun fazhan xiehui; Telecommunication Development Association of Republic of China), 農業試驗所 (nongye shiyansuo; Agriculture Research Institute), 開拓文教基金會 (kaituo wenjiaojijinghui; Kaituo Cultural and Educational Foundation), 國際翻譯社 (guoji fanyishe; International Translation Agency).

To resolve this problem, a refined method was used to allow these words to serve as the name parts of organization names. The performance of the refined method is shown in Table 1(c). With this heuristic rule, the precision was 65.15% and the recall was

74.79% in the anchor part. Appendix B presents some extracted examples in the anchor part. The refined method achieved a precision rate of 61.38% and a recall rate of 68.74% for the content part.

In the content part of the 703 web pages, there were 1,732 proper names and 3,029 organization names. Only one of these proper nouns or none was the owner of the web page. That is, at least 85.23% of these names were unrelated to the owners of the web pages. Totally, 64 E-mail addresses and 16 HTTP URLs were extracted in the non-anchor part. Because the patterns of the E-mail addresses and HTTP URLs were well-formed, all of them were found. These addresses and URLs were related to none or one of 6,735 proper nouns (3,343 personal names and 3,392 organization names). With the mapping heuristics, 18 E-mail addresses were assigned to the correct personal names or organization names; 5 E-mail addresses were assigned incorrectly; and the others were not assigned. The mapping algorithm achieved an accuracy rate of 78.26%. We found that the Pinyin spelling similarity provided a very good criterion to relate the E-mail addresses to the proper nouns, even when they were not the nearest pairs. Some experimental data and results are shown in Appendix C.

Table 2 summarizes the overall results of information extraction for proper nouns. 97.52% of the information was extracted from the anchor set. The number of home page URLs and E-mail addresses extracted in the content part was much smaller than that in the anchor part. This reflects the characteristics of web pages. When designing web pages, people often include URLs and E-mail addresses within the linking anchors for users' navigation instead of giving the information in the content. Because the HTML anchor tags explicitly give the information about the linking text, the overall performance will depend on the identification of proper nouns in the anchors of the web pages.

Table 2. *The Overall Results of Information Extraction for People Finding*

Information Extraction	# (no. of correct items)	% (Percentage)
From Anchor Set	747 (189 for people and 558 for organizations)	97.52%
From Content Set	19	2.48%
Total	766	100%

The major errors resulted from conjunctions and compounds in the organization names. For complex proper names, the correct boundaries were not determined in the identification task. Some examples are shown in the following. In the string ' 台大建築與城鄉研究所 ' (Taida Jianzhu yu Chengxiang yanjiusuo; Graduate Institute of Building and Planning), an organization name ' 城鄉研究所 ' (Chengxiang yanjiusuo) was identified with an incorrect left boundary because of the conjunction ' 與 ' (yu; and).

<A href="http://www.bp.ntu.edu.tw/">台大建築與城鄉研究所 / Graduate Institute of Building and Planning </A>
Oname: 城鄉研究所
公立大學暨獨立學院 / Public University and College
Oname: 公立大學
公立大學暨獨立學院 / Public University and College
Oname: 獨立學院
長庚醫學暨工程學院 / Chang Gung College of Medicine and Technology
Oname: 工程學院
教育網路中心 / Educational Network Center
Oname: 網路中心
華梵人文科技學院 / Huafan College of Humanities and Technolgy
Oname: 科技學院

In the content part of the web pages, the system produced some incorrect personal names. For example, the personal name ' 魏晉南 ' (Wei Jin Nan) was incorrectly identified in the famous dynasty ' 魏晉南北朝 ' (Wei Jin Northern and Southern Dynasties), because ' 魏 ' (Wei) is a frequently-used surname and ' 晉南 ' (Jin Nan) is like a name. Further, some famous historical books and names of years were very similar to the personal names. On the other hand, some famous ancient personal names could not be identified, because the name parts of these names were rarely used in the training corpus of contemporary personal names. In addition, nicknames and some transliterated names were missed. This is because a nickname does not lead with a surname, and most of the characters used in Japanese names are different from those in transliterated English names. Some examples are listed below.

(Incorrect Identification)

- Famous dynasty:
  魏晉南 (魏晉南北朝 ; Wei Jin Northern and Southern Dynasties),
  魏晉隋 (Wei Jin & Sui Dynasties), 魏晉 (Wei & Jin Dynasties),
  隋唐 (Sui & Tang Dynasties)
- History books:
  史傳 (Shizhuan), 白書 (Baishu), 古史 (Gushi; Ancient History)
- Year: 丁巳 (Ding Si)
- Transliterated Japanese Names: 山根 (山根幸夫 ; Yamane Yukio)

(Miss)

- Ancient People:
  司馬遷 (simaqian), 顧炎武 (guyanwu), 劉知幾 (liuzhiji), 胡適 (Hushi), 逯耀東 (luyaodong), 郭沫若 (guomuruo)
- Nicknames:
  小安安 (xiao anan), 小桂子 (xiao guizi), 阿勳 (a xun), 阿賢 (a xian), 潔潔 (jiejie), 雄雄 (xiongxiong)

●Transliterated Japanese Names:
雄川一郎 (Okawa Ichiro), 小林直樹 (Kobayashi Naoki)

To increase the coverage of the dictionary can reduce the error rates in name recognition. For example, famous dynasties, history books, and famous ancient personal names should be added to the lexicon. Furthermore, nicknames and transliterated names (e.g., Japanese names) should be investigated further.

7. Concluding Remarks

This paper has proposed a computer-aided information extraction method to construct white pages for Internet/Intranet users or to build databases for finding people and organizations on the Internet. The traditional approach used by current search engines indexes proper nouns with incorrect URLs of web pages in the task of finding people and organizations. In our system, proper nouns are identified using some heuristic rules and the corpus-based analysis method of natural language processing. Considering the semantics of content and HTML tags, these proper nouns and their related information are extracted from web pages. Using identification of proper nouns, the number of indexing terms on a web page using the proposed method is smaller than that using search engines. Finding people and organizations in the database of the extracted results is more precise than in the current search engines. The results here show that much interesting information can be automatically extracted from the WWW. However, complete identification of conjunctions and compounds in organization names needs further investigation. Furthermore, other types of information, e.g., addresses, phone numbers, and so on, will be considered in the future.

References

Bian, G.W. and Chen, H.H. "An MT Meta-Server for Information Retrieval on WWW", *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, Palo Alto, California, USA, March, 1997, pp.10-16.

Boguraev, B. and Pustejovsky, J. *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA, USA., 1996.

Chang, J.S., et al. "Large-Corpus-Based Methods for Chinese Personal Name Recognition", *Journal of Chinese Information Processin*g 6.3 (1992): pp. 7-15.

Chen, H.H and Lee, J.C. "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 15th International Conference on Computational Linguistics*, 1996, pp. 222-229.

Church, K.W. and Hanks, P. "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics* 16.1 (1990): pp. 22-29.

CL, "Special Issues on Using Large Corpora," *Computational Linguistics* 19. 1-2, 1993.

Davis, M.W. and Ogden, W.C. "Implementing Cross-Language Text Retrieval Systems for Large-scale Text Collections and the World Wide Web." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 9-17.

Etzioni , "Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web", URL ftp://ftp.cs.washington.edu/pub/etzioni/softbots/a96.ps.gz, *Proceeding of AAAI-96*, 1996.

Gachot, D.A.; Lange, E. and Yang, J. "The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Multilingual Information Retrieval." *Proceedings of Workshop on Cross-Linguistic Information Retrieval,* 1996, pp. 44-54.

Hayashi, Y.; Kikui, G. and Susaki, S. "TITAN: A Cross-linguistic Search Engine for the WWW." *Working Notes of the AAAI-97 Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997, pp. 58-65.

HTML, *HyperText Markup Language*, URL http://www.w3.org/pub/WWW/Markup, 1996.

Lu, Suping, "A Study on the Chinese Romanization Standard in Libraries," *Cataloging and Classification Quarterly* 21 (1995):81-97.

Mani, I., et al. "Identifying Unknown Proper Names in Newswire Text," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44-54.

McDonald, D. "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 32-43.

Paik, W., et al."Categorization and Standardizing Proper Nouns for Efficient Information Retrieval," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 154-160.

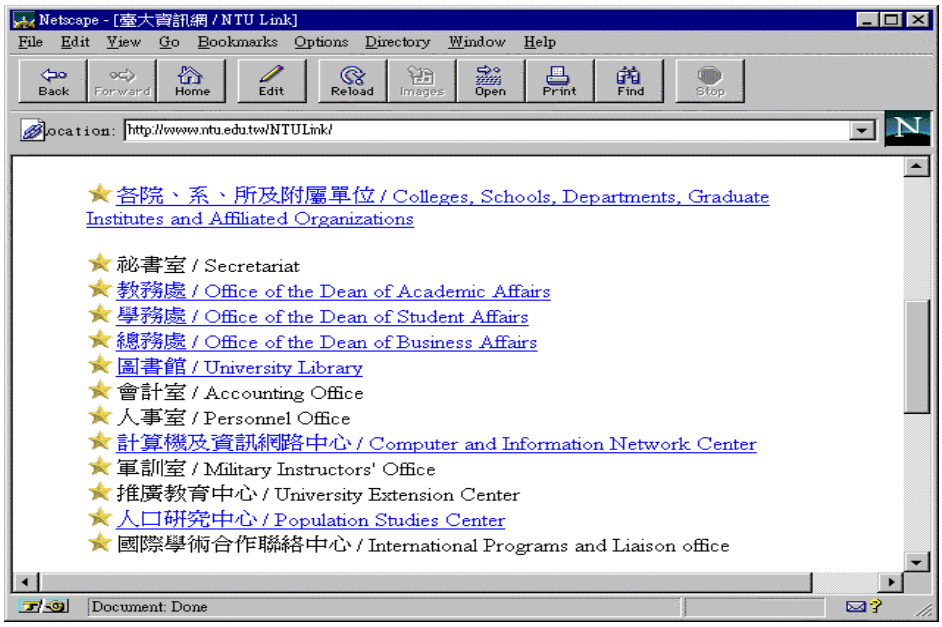
Sproat, R., et al."A Stochastic Finite-State Word-Segmentation Algorithm for Chinese", *Proceeding of 32nd Annual Meeting of ACL, New Mexico*, 1994, pp. 66-73.

Wang, L.J; Li, W.C. and Chang, C.H. "Recognizing Unregistered Names for Mandarin Word Identification", *Proceeding of 14th COLING*, Nantes, 1992, pp. 1239-1243.

Appendix A. Hierarchical Features of Home Pages

(1) Home page of National Taiwan University

(2) Home Page from NTU Link



Appendix B. Some Experimental Results in the Anchor Part

In the following, Oname and Pname denote the extracted organization names and personal names, respectively.

[Organization-School (Oname)]

國立臺灣大學 / National Taiwan University Oname: 國立臺灣大學
國立政治大學 / National Chengchi University Oname: 國立政治大學
國立清華大學 / National Tsing Hua University Oname: 國立清華大學
國立交通大學 / National Chiao Tung University Oname: 國立交通大學
國立臺灣師範大學 / National Taiwan Normal University Oname: 國立臺灣師範大學
國立中央大學 / National Central University Oname: 國立中央大學
國立中山大學 / National Sun Yat-sen University Oname: 國立中山大學
國立成功大學 / National Cheng Kung University Oname: 國立成功大學
國立中正大學 / National Chung Cheng University Oname: 國立中正大學
國立陽明大學 / National Yang Ming University Oname: 國立陽明大學
國立東華大學 / National Dong Hwa University Oname: 國立東華大學
國立臺灣海洋大學 / National Taiwan Ocean University Oname: 國立臺灣海洋大學
國立暨南國際大學 / National Chi-Nan University Oname: 國立暨南國際大學
中央警察大學 / Central Police University Oname: 警察大學
國立台北師範學院 / National Taipei Teachers College Oname: 國立台北師範學院
台北市立師範學院 / Taipei Municipal Teachers College Oname: 台北市立師範學院
國立藝術學院 / National Institute of the Arts Oname: 國立藝術學院
國立台北護理醫學院 / National Taipei College of Nursing Oname: 國立台北護理醫學院
國立台灣工業技術學院 / National Taiwan Institute of Technology Oname: 國立台灣工業技術學院
<A HREF="http://www.princeton.edu/index.html">普林斯頓大學</A> Oname: 普林斯頓大學
慈濟醫學院 / Tzu Chi College of Medicine Oname: 慈濟醫學院
朝陽技術學院 / Chaoyang Institute of Technology Oname: 朝陽技術學院
元智工學院 / Yuan-Ze Institute of Technology Oname: 元智工學院
高雄工學院 / Kaohsiung Polytechnic Institute Oname: 高雄工學院
中華工學院 / Chung-Hua Polytechnic Institute Oname: 中華工學院
大葉工學院 / Da-Yeh Institute of Technology Oname: 大葉工學院
國立臺北商業專科學校 / National Taipei College of Business
Oname: 國立臺北商業專科學校
國立臺中商業專科學校 / National Taichung Institute of Commerce
Oname: 國立臺中商業專科學校
國立屏東商業專科學校 / National Pingtung Institute of Commerce
Oname: 國立屏東商業專科學校
國立嘉義農業專科學校 / National Chia-Yi Institute of Agriculture
Oname: 國立嘉義農業專科學校
國立宜蘭農工專科學校 / National Ilan Institute of Agriculture and Technology
Oname: 宜蘭農工專科學校
國立高雄工商專科學校 / National Kaohsiung Institute of Technology
Oname: 國立高雄工商專科學校
國立勤益工商專科學校 / National Chinyi Institute of Technology
Oname: 國立勤益工商專科學校
國立聯合工商專科學校 / National Lien-Ho College of Technology and Commerce
Oname: 國立聯合工商專科學校
國立雲林工業專科學校 / National Yunlin Polytechnic Institute
Oname: 國立雲林工業專科學校
國立高雄餐旅管理專科學校 / National Kaohsiung Hospitality College
Oname: 國立高雄餐旅管理專科學校
國立台灣體育專科學校 / National Taiwan College of Physical Education
Oname: 國立台灣體育專科學校
臺南家政專科學校 / Tainan College of Home Economics Oname: 臺南家政專科學校
佛教慈濟護理專科學校 / Buddhist Tz'u Chi Junior College of Nursing
Oname: 慈濟護理專科學校
健行工商專校 / Chien Hsien Institute of Technology and Commerce Oname: 健行工商
萬能工商專科學校 / VanNung Institute of Technology Oname: 萬能工商專科學校
南亞工商專科學校 / Nanya Junior College Oname: 南亞工商專科學校
龍華工商專科學校 / Lunghwa Junior College of Technology and Commerce
Oname: 龍華工商專科學校

明新工商專校 / Ming Hsin Institute of Technology Oname: 明新工商
大華工商專科學校 / Ta Hua College of Technology and Commerce
Oname: 大華工商專科學校
親民工商專科學校 / Chin Min College of Technology and Commerce
Oname: 親民工商專科學校
樹德工商專科學校 / Shu Teh Junior College of Technology Oname: 樹德工商專科學校
中州工商專校 / Chung Chou Junior College of Technology and Commerce
Oname: 中州工商專校
建國工商專科學校 / Chienkuo Junior College of Technology Oname: 建國工商專科學校
吳鳳工商專科學校 / Wu-Feng Junior College of Technology and Commerce
Oname: 吳鳳工商專科學校
南台工商專科學校 / Nan Tai College of Technology and Commerce
Oname: 南台工商專科學校

[Organization-Club (Oname)]

台大佳韻音樂社 Oname: 佳韻音樂社
台大鋼琴社 Oname: 鋼琴社
杏林合唱團 Oname: 杏林合唱團
杏林弦樂團 Oname: 弦樂團
基克工作室 Oname: 基克工作室

[Organization-Government (Oname)]

網路博覽會 中華民國館 / Pavilion of Taiwan, R.O.C. Oname: 中華民國館
<A HREF="http://expo96.org.tw/Welcome_c.html">中華民國館</A> Oname: 中華民國館
<A HREF="http://www.motc.gov.tw/Welcome_c.html">交通館</A> Oname: 交通館
<A HREF="http://www.nmns.edu.tw/">國 立 自 然 科 學 博 物 館 </A> Oname: 國立自然科學博物館
台 北 市 立 動 物 園 Oname: 台北市立動物園
<A HREF="http://192.192.14.202/welcome.htm">國立中正文化中心</A> Oname: 國立中正文化中心
<A HREF="http://crab.ccl.itri.org.tw/cgi/m_normal">國家圖書館遠距圖書服務系統</A> Oname: 國家圖書館

[Personal name (Pname)]

"http://dodger.ee.ntu.edu.tw/~lswang/">王立三的 HomePage / Li-San Wang's Homepage Pname: 王立三
"http://www.csie.ntu.edu.tw/~jcwang/index.cgi">王家俊 / John's House Pname: 王家俊
"http://med.mc.ntu.edu.tw/~shouzen/">生命的照顧 － 范守仁醫師 / Life Care - Fan's Home Pname: 范守仁
"http://king.cc.ntu.edu.tw/~d0701021/hgt/">何子之網頁 Pname: 何子
"http://www.ee.ntu.edu.tw/~b82070/">杜立群 Pname: 杜立群
"http://nlg3.csie.ntu.edu.tw/group/gwbian.html">邊國維的網頁 Pname: 邊國維
"http://osil.csie.ntu.edu.tw/~chwu/">吳俊興 Pname: 吳俊興
"http://king.cc.ntu.edu.tw/~b3401111/">吳振漢的窩 / Wilfred's HomePage Pname: 吳振漢
"http://king.cc.ntu.edu.tw/~b3502118/">林育德（AirL)的遊園地 Pname: 林育德
"http://king.cc.ntu.edu.tw/~b2504049/">林欣蔚 / CELHW Pname: 林欣蔚
"http://ipmc.ee.ntu.edu.tw/~sclin/">林信成的 W3 小棧 Pname: 林信成
"http://king.cc.ntu.edu.tw/~b2501109/welcome.htm">依客那米克斯傳說—勇者耀耀之章 Pname: 那米克斯
"http://140.112.19.6:8000/">阿哲的夢幻天地 Pname: 阿哲
"http://med.mc.ntu.edu.tw/~green/">林錦鴻 - 電腦玩家，網路流民，婦產科醫師 Pname: 林錦鴻
"http://king.cc.ntu.edu.tw/~b2501127/">唐唐的世界 Pname: 唐唐
"http://king.cc.ntu.edu.tw/~b2603230/">張正宜-不來不可的好地方 / TOM's Home Pname: 張正宜
"http://sun.gcc.ntu.edu.tw/Huang/">黃兆談 Pname: 黃兆談
"http://king.cc.ntu.edu.tw/~r5241206/">魚兒的小鎮－林康捷的 Homepage Pname: 林康捷
"http://king.cc.ntu.edu.tw/~b3503015/">陳紀光 / HomePage of Chen Chi-kuang Pname: 陳紀光
"http://cml19.csie.ntu.edu.tw/~robin/">陳炳宇 / Robin's Workgroup Pname: 陳炳宇
"http://med.mc.ntu.edu.tw/~b9401011/">郭昇彥的烘焙機 Pname: 郭昇彥

Appendix C. Some Mapping Results in the Content Part

In the following, Oname and Pname denote the extracted organization names and personal names, respectively. The number indicates the token no. of the information in the web pages.

[Some Extracted Data in Content Sets before Mapping]

Oname: 資訊新館 63
E-Mail: root@csman.csie.ntu.edu.tw 69

Oname: 土木館 81
E-Mail: root@ce.ntu.edu.tw 82

Pname: 蔡博文 108
Oname: 地理系館 109
E-Mail: tsaibw@ccms.ntu.edu.tw 112

Pname: 丘台生 122
Oname: 漁科館 123
E-Mail: tschiu@ccms.ntu.edu.tw 124

Pname: 陳膺州 146
E-Mail: ingchen@chem60.ch.ntu.edu.tw 152

Pname: 張震東 155
E-Mail: gdchang@ccms.ntu.edu.tw 160

Pname: 黃靜美 171
E-Mail: mei@ccms.ntu.edu.tw 175

Pname: 林翰彥 178
Oname: 森林館 179
E-Mail: wenliang@ccms.ntu.edu.tw 180

Pname: 蘇明道 184
Oname: 農工館 185
E-Mail: sumd@ccms.ntu.edu.tw 186

Pname: 王友俊 382
E-Mail: wangecaa@ccms.ntu.edu.tw 387

Pname: 周伯戡 389
E-Mail: pkchou@ccms.ntu.edu.tw 391

Pname: 游張松 250
E-Mail: yucs@ccms.ntu.edu.tw 254

Pname: 曾珀雯 270
Pname: 徐信權 272
E-Mail: popo@ccms.ntu.edu.tw 276
E-Mail: kevins@ccms.ntu.edu.tw 277

[Some Mapping Results in Content Sets]

E-Mail: root@csman.csie.ntu.edu.tw
E-Mail: focus@www.ntu.edu.tw
E-Mail: news@www.ntu.edu.tw
E-Mail: campus@www.ntu.edu.tw
E-Mail: tsaibw@ccms.ntu.edu.tw
E-Mail: tschiu@ccms.ntu.edu.tw
E-Mail: ingchen@chem60.ch.ntu.edu.tw
E-Mail: yucs@ccms.ntu.edu.tw
E-Mail: hlee@cc.ntu.edu.tw
E-Mail: popo@ccms.ntu.edu.tw
E-Mail: kevins@ccms.ntu.edu.tw
http: http://www.ntu.edu.tw/forest/R17.html

Oname: 資訊新館
Oname: 焦點新聞
Oname: 網路新聞
Oname: 校園新聞
Pname: 蔡博文
Pname: 丘台生
Pname: 陳膺州
Pname: 游張松
Pname: 李賢輝
Pname: 曾珀雯
Pname: 徐信權
Oname: 國立臺灣大學森林學系暨研究所