Proper Name Extraction from Web Pages for Finding People in Internet

Hsin-Hsi Chen and Guo-Wei Bian

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.

E-mail: hh_chen@csie.ntu.edu.tw; gwbian@nlg.csie.ntu.edu.tw

Abstract

This paper proposes a method to extract proper names and their associated information from Web pages for Internet/Intranet users automatically. It extracts information from World Wide Web documents, including proper nouns, E-mail addresses and home page URLs, and finds the relationship among these data. Natural language processing techniques are employed to identify and classify proper nouns, which are usually unknown words. Different kinds of clues such as spelling method, adjacency principle and HTML tags are used to relate proper nouns to their corresponding E-mail and/or URL. With the mapping schemes, the extracted information is more accurate than the results from the traditional searching engines. The results can be used as the database of the services for finding people and organizations in Internet. Such searching services are very useful for human communication and dissemination of information.

1. Introduction

With the rapid growth of Internet in recent years, World Wide Web (WWW) becomes a very large knowledge source nowadays. Much information is disseminated through the giant media. One of the problems in the large cyber space is: it is very difficult to know where an entity, which is a concrete object that can send and receive information, is. For communication purpose, we usually want to know a person's or a company's E-mail address, or his/her home page URL. Yellow pages, which are E-mail directory or URL directory in this case, can help to find the related information. Because WWW is a very large database and is created dynamically, it is hard to set up such a kind of yellow pages manually. In the other way, the current searching engines just index the contents of the Web page with their URL. Supposing a page contains many proper names, the searching 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. For the need of intelligent processing, the precision of the searching engines is too low for such a task of finding people.

Hopefully, very large portion of WWW is composed of natural language documents that can be regarded as a text corpus. Corpus analysis techniques in natural language processing (CL, 1993) can be employed to extract knowledge from WWW. And using the semantics of the content and HTML tags, some mapping schemes are proposed to relate the knowledge with the URL of Web page.

This paper will propose a method to construct yellow 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 relationship among these data. The problems to be tackled are shown as follows:

(1) Proper nouns, which are always unknown words, have to be identified and classified from WWW corpus. Those proper nouns that denote organizations are usually hierarchical. Such kinds of relationships must be distinguished.

(2) There may be more than one proper noun, more than one E-mail, 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 language models proposed in this paper are experimented on Taiwan WWW home pages. 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 shows the overview of our yellow 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 mapping algorithms between proper nouns and the related information. Section 6 discusses the experiments, and section 7 concludes some remarks.

2. WWW Documents

The first step in constructing yellow pages is to know where the proper nouns, E-mail addresses and URLs are in WWW documents. WWW documents are different from the traditional text corpus in that they are HTML (HyperText Markup Language) files. The tagging information provides some clues, but it also introduces some noises. How to use the information is a very important issue in applications on Internet, e.g., cross-language information retrieval (Bian and Chen, 1997). In plain text, each sentence always has a sentence terminator such as full stop, question mark and 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 of 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)

Besides, some punctuation symbols like '|' and ':' have the same effects. In contrast to the above sentence delimiters, the font style elements may introduce noises. 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. Thus these tags should be treated as meta-information and hidden from processing.

The links denoted by anchors (A) in the WWW documents are one of the possible sources of proper nouns and the related information. Some WWW documents shown in Appendix A demonstrate 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 NTU ('國立台灣大學'), which is an organization name. Underline shows a link to other home pages. The second example follows from NTU Link. The interesting entities are Offices of Academic Affairs ('教務處'), of Student Affairs ('學務處'), 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') are not constructed now, so that they are not listed in the final yellow 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. A link in the HTML file may be represented as follows:

<a href="argument"> text </a>

When “text” is a proper noun, its home page URL may be described by “argument”. Consider an example in the NTU Link home page. The link of Office of the Dean of Academic Affairs (‘教務處’) is shown below:

```
<a href="/Campus/announce/index.html#academic">教務處
            / Office of the Dean of Academic Affairs</a>
```

If the proper noun and its URL are put into yellow pages directly, this entry may be ambiguous. This is because many universities have the similar organization. Therefore we should keep the hierarchical path of the Web pages to disambiguate the meaning of a proper noun. Further, the relative URLs need to be modified as the absolute ones for keeping the complete URL information.

Besides the link field, proper nouns may appear in other positions in a WWW document. To deal with these objects is more complex. An additional algorithm is needed to associate URLs and E-mail addresses with suitable proper nouns. Different kinds of clues such as spelling method, adjacency principle and HTML tags (e.g., title, headings, address, font style elements) are employed.

3. System Overview

We periodically collect the home pages from Internet/Intranet by a spider. The yellow page constructor first analyzes these HTML files. The 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 anchor parts is placed in the anchor set (AS). Other information, i.e., that appears in non-anchor parts, is placed in one of the content sets (CSes) for the different types of information. In current implementation, there are three content sets - say, 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. For the content sets, a mapping algorithm (see Section 5) associates URLs and/or E-mail addresses with a suitable proper noun. Algorithm 1 shows the information extraction part of the yellow page constructor.

Algorithm 1. Information Extraction

```
Input:    An HTML file or a plain text

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. For each processing unit
             3.1 [Extracting the Anchor Information]
                 for each anchor (<a href=" protocol://host/path ">Text</a>)
                 {  Identify and classify proper noun (PN) within Text.
                    if PN exists, add the tuple (PN, protocol://host/path) to the
                       Anchor Set (AS)
                 }
             3.2 [Extracting the Content Information]
                 3.2.1 Identify and classify proper nouns (PNs)
                       if found
                       { add PN to CS_Proper-Noun with the following attributes:
                            position information (token_no) and associated HTML
                            meta information (<TITLE>, <Hn>, <Address>,
                            <Bold>, <Font> and <Italic>)
                       }
                 3.2.2 Extract different types of information with token_no, and
                       add to the corresponding Content Sets (CSes).

          4. End
```

4. Identification of Proper Nouns

Proper nouns which are not collected in lexicons are major unknown words in natural language texts. Several methods (Boguraev and Pustejovsky, 1996; Chen and Lee, 1996; Mani, *et al.*, 1993; McDonald, 1993; Paik, *et al.*, 1993) have been proposed to identify proper nouns. Of these, Chen and Lee (1996) present various strategies to identify and classify three types of proper nouns in Chinese texts, i.e., Chinese personal names, Chinese transliterated personal names and organization names. In large-scale experiments, the average precision rate is 88.04% and the average recall rate is 92.56% for the identification of Chinese personal names. The average precision rate and the average recall rate for the identification of organization names are 61.79% and 54.50%, respectively. We follow this work on the extraction of personal names and organization names from Taiwan Web pages.

4.1 Identification of Personal Names

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

Model (a) Single character

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

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

Model (b) Two characters

$$(3)\ \frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold4$$

Model (c) Two surnames together

$$(4)\ \frac{\#C_{11}}{\&C_{11}}\times\frac{\#C_{12}}{\&C_{12}}\times\frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold5$$

$$(5)\ \frac{\#C_{11}}{\&C_{11}}\times\frac{\#C_{12}}{\&C_{12}} > Threshold6 \text{ and } \frac{\#C_2}{\&C_2}\times\frac{\#C_3}{\&C_3} > Threshold7$$

For different types of surnames, different models are adopted. Because the surnames with two characters are always surnames, Model (b) neglects the score of surname part. Models

(a) and (c) have two score functions. They solve the problem of very high score of surnames. The above three models can be extended to the single-character names. When a candidate cannot pass the thresholds, its last character is cut off and the remaining string is tried again. Thresholds are trained from a large-scale Chinese name corpus of 219,738 Chinese personal names. We let 99% of the training data pass the thresholds.

Text provides many useful clues from three different levels - say, character, sentence and paragraph levels. The baseline model belongs to the character level. Titles, mutual information and punctuation marks come from sentence level. When a title such as "President" appears before (after) a string, it is probably a personal name. Mutual information (Church and Hanks, 1990), which provides a measure of word association, is employed to tell the difference between a name and a content word. We check the string which can serve as a name or a content word with its surrounding words. When they have a strong relationship, it has high probability to be a content word rather than a name. The punctuation marks play an important role in identification. Personal names usually appear at the head or the tail of a sentence. The last clue is the paragraph information. A personal name may appear more than once in a paragraph. We use cache to store the identified candidates, and reset cache when next paragraph is considered.

4.2 Organization Names

The structures of organization names are more complex than those 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 keyword is an important clue to extract the organization names. However there are still several difficult problems. First, a keyword is usually a common content word. It is not easy to tell their difference. Second, a keyword may appear in an abbreviated form, or even be omitted completely. Third, some organization names are very long, so it is hard to decide the left boundary.

This paper only touches on the third problem. Keywords, which are good indicators, play the similar role of surnames. They show not only the possibility of an occurrence of an organization name, but also its right boundary. Prefix is a good marker for possible left boundary. Parts of speech such as transitive verbs, adjectives, numerals and classifiers are also useful to determine the left boundary. The name part of an organization cannot beyond these critical parts of speech. Because a tagger is not involved before identification, the part of speech of a word is determined wholly by its lexical probability. Finally, the formulation of the name part of an organization name is considered. If the word preceding a keyword is a place name or a personal name, it forms the name part of an organization. Otherwise, we use the word association model to determine the left boundary. The postulation is: the words to compose a name part usually have strong relationship. The mutual information mentioned in the last section is also used to measure the relationship of two words.

5. A Mapping Algorithm

Algorithm 2 is a mapping algorithm that relates URLs and/or E-mail addresses to the proper nouns. A score function that considers spelling method, adjacency principle and HTML tags is used to determine the relationship among proper nouns and the related information.

```
Algorithm 2. Information Mapping
Input:      Three Contents Sets (CSs)
            A Threshold and a Window_Size of context
Output:     A Mapping Set (MS)
Function:   Mapping CS_E-mail (CS_HTTP) with CS_Proper-Name
Method:     1. Set MS to an empty set.
            2. For each CS 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 Score(Info, PN) > Threshold
                     {    add (Info, PN) into MS
                     }
                  }
               }
          3. End
```

The ranking function is defined as follows:

$$Score(\text{Info, PN}) =$$

$$\left(\frac{HTML\_SCORE(\text{PN})+1}{\text{abs(Info.token\_no - PN.token\_no)}} + \frac{Title(PN)}{\text{Total\_tokens - Info.token\_no+1}}\right)$$

$$+ Pinyin\_Similarity(\text{PN, Info})*E\text{-}mail(\text{Info}) *1.2$$

$$HTML\_SCORE(\text{PN}) =$$

$$Title(PN) + Heading(PN) + Address(PN) + Bold(PN) + Font(PN) + Italic(PN)$$

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

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 user-id in E-mail address is often transliterated from Chinese name, the similarity has the highest priority than the other cues. We often adopt a Pinyin system (Lu, 1995) to transliterate Chinese name. The Pinyin Similarity is defined as follows:

   $$Pinyin\_Similarity(\text{PN, E-mail}) =$$

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

   For example, the Pinyin transliteration of "邊國維" is "Bian Guo Wei".

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

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

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

2. **Adjacency Principle.** Proper nouns and the related information are often near. The distance between Info and PN is measured in terms of the number of intervened tokens. Recall that we assign each object a unique token number. The closer pair has a larger score.

3. **HTML Tags.** The proper nouns (PNs) that appear in Title (<Title>) / Heading(<Hn>…</Hn>) / Address, or are described by the font style (Bold,

Italic and Font tag elements) are given larger weight for ranking than other normal proper nouns.

6. Experiments

In our initial experiments, total 703 home pages are collected from our campus NTU Web (http://www.ntu.edu.tw/). The collected pages are classified into an anchor set and a content (non-anchor) set. Then, the personal names and organization names are corrected by human for evaluation. The window size (Window_Size) of context is 6 and the score threshold (Threshold) is 0.2 for the mapping algorithm. Table 1 shows the results of identification in both sets and the mapping result in the content set. Appendix B and C demonstrate some extracted examples.

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

(a) Identification of Proper Nouns in the Anchor Set

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

(b) Identification of Proper Nouns 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%

(c) The Mapping Result in the Content Set

Table 1 The Results of Identification and Information Mapping.

In anchor part, there are 6204 linking items. Of these, the numbers of personal names and organization names are 255 and 746, respectively. That is, 83.87% that are irrelevant should be screened for the task of finding people. The precision and the recall are 82.89% and 74.12% for the identification of personal names. But the precision and the recall for the identification of organization names are much lower than those in our previous work. The major errors result from the conjunctions and compounds of the organization names. For these complex proper names, the correct boundaries are not determined in identification task. Some examples of errors are shown in the following.

```
<A href="http://www.bp.ntu.edu.tw/">台大建築與城鄉研究所</A>
          Oname: 城鄉研究所
<a href="http://jojo.ntu.edu.tw/TANet/public.html">公立大學暨獨立學院 / Public University and College</a>
          Oname: 公立大學
<a href="http://jojo.ntu.edu.tw/TANet/public.html">公立大學暨獨立學院 / Public University and College</a>
          Oname: 獨立學院
<a href="http://linux1.cgu.edu.tw/">長庚醫學暨工程學院 / Chang Gung College of Medicine and Technology</a>
          Oname: 工程學院
<a href="http://jojo.ntu.edu.tw/TANet/edu.html">教育網路中心 / Educational Network Center</a>
          Oname: 網路中心
<a href="http://www.hcht.edu.tw/">華梵人文科技學院 / Huafan College of Humanities and Technolgy</a>
          Oname: 科技學院
```

In the other way, there are 1732 proper names and 3029 organization names listed in the content part of the 703 Web pages. Only one of these proper nouns or none is the owner of one page. At least, 85.23% of these names are irrelevant. The current searching engine will index all the proper nouns with their URLs. This is the reason why the precision of the searching engines is too low for such a task of finding people.

Totally, there are 64 E-mail addresses and 16 HTTP URLs extracted in the non-anchor part. With the mapping heuristics, 18 E-mail address are assigned the correct personal names or organization names; 5 E-mail addresses are assigned incorrectly; and the others have no associated ones. The mapping algorithm achieves 78.26% accuracy to relate the information with the proper nouns. We found the spelling Pinyin similarity provides very good heuristics to relate the E-mail addresses to the proper nouns, even they are not the nearest pairs. Some experimental data and results are shown in Appendix C.

7. Concluding Remarks

This paper proposes a computer-aided information extraction method to construct yellow pages for Internet/Intranet users or to build the database of the services for finding people and organizations in Internet. The results show much interesting information can be extracted from WWW. However, the complete identification for the conjunction and compound of the organization names need further investigations in future works. Other types of information, e.g., addresses, phone numbers, and so on, will be considered in the next step. Besides, the hierarchical relationship should be tackled to set up complete yellow pages.

References

Bian, G.W. and Chen, H.H. (1997). "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. (1996). *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA, USA., 1996.

Chen, H.H and Lee, J.C. (1996) "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. (1990) "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16, No. 1, 1990, pp. 22-29.

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

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

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

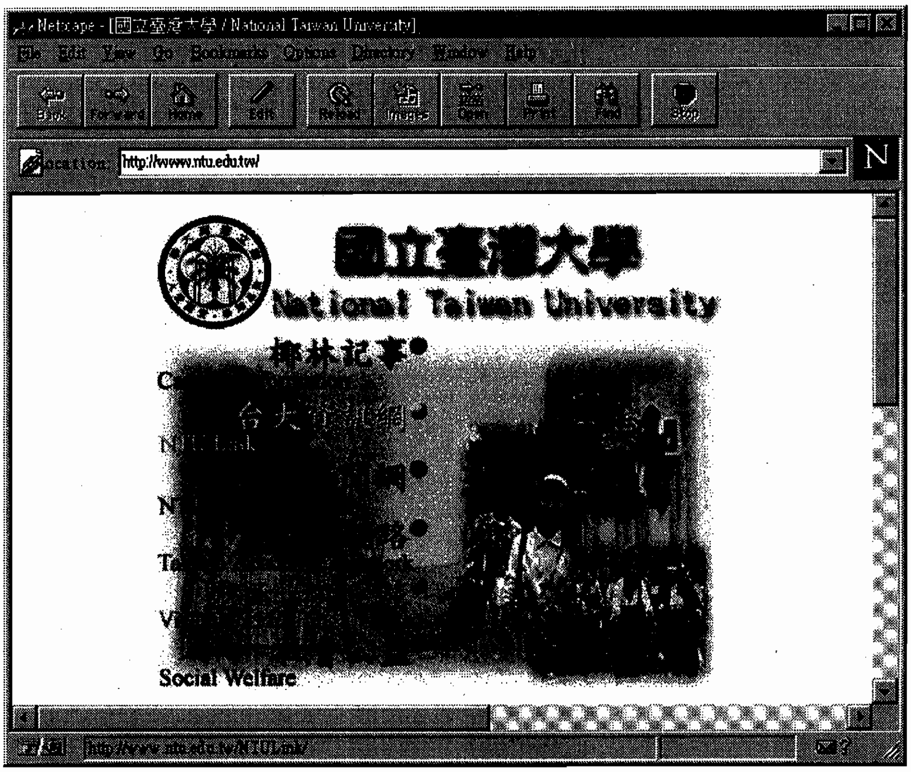
Mani, I., *et al.* (1993) "Identifying Unknown Proper Names in Newswire Text," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44-54.

McDonald, D. (1993) "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.

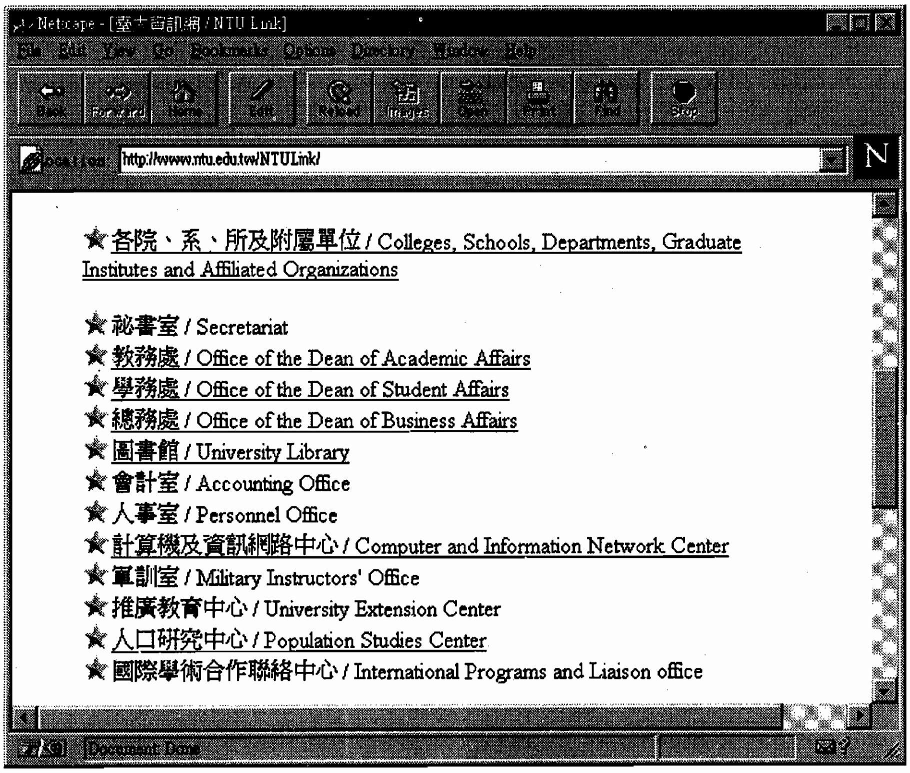
Paik, W., *et al.* (1993) "Categorization and Standardizing Proper Nouns for Efficient Information Retrieval," *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 154-160.

Appendix A. Hierarchical Features of Home Pages

(1) Home page of National Taiwan University


Netscape - [國立臺灣大學 / National Taiwan University]
File Edit View Go Bookmarks Options Directory Window Help
Back
Forward
Home
Edit
Reload
Images
Open
Print
Find
Stop
Location: http://www.ntu.edu.tw/
N
國立臺灣大學
National Taiwan University
椰林記事
台大資訊網
NTU Link
Social Welfare
http://www.ntu.edu.tw/NTULink/

(2) Home Page from NTU Link


Netscape - [臺大資訊網 / NTU Link]
File Edit View Go Bookmarks Options Directory Window Help
Back
Forward
Home
Edit
Reload
Images
Open
Print
Find
Stop
Location: http://www.ntu.edu.tw/NTULink/
N
Document: Done

★各院、系、所及附屬單位 / Colleges, Schools, Departments, Graduate Institutes and Affiliated Organizations

★祕書室 / Secretariat
★教務處 / Office of the Dean of Academic Affairs
★學務處 / Office of the Dean of Student Affairs
★總務處 / Office of the Dean of Business Affairs
★圖書館 / University Library
★會計室 / Accounting Office
★人事室 / Personnel Office
★計算機及資訊網路中心 / Computer and Information Network Center
★軍訓室 / Military Instructors' Office
★推廣教育中心 / University Extension Center
★人口研究中心 / Population Studies Center
★國際學術合作聯絡中心 / International Programs and Liaison office

Appendix B. Some Experimental Results in Anchor Part

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

[Organization-School (Oname)]

```
<a href="http://www.ntu.edu.tw/">國立臺灣大學 / National Taiwan University</a>                Oname: 國立臺灣大學
<a href="http://www.nccu.edu.tw/">國立政治大學 / National Chengchi University</a>             Oname: 國立政治大學
<a href="http://www.nthu.edu.tw/">國立清華大學 / National Tsing Hua University</a>            Oname: 國立清華大學
<a href="http://www.nctu.edu.tw/">國立交通大學 / National Chiao Tung University</a>           Oname: 國立交通大學
<a href="http://www.ntnu.edu.tw/">國立臺灣師範大學 / National Taiwan Normal University</a>    Oname: 國立臺灣師範大學
<a href="http://www.ncu.edu.tw/">國立中央大學 / National Central University</a>               Oname: 國立中央大學
<a href="http://www.nsysu.edu.tw/">國立中山大學 / National Sun Yat-sen University</a>         Oname: 國立中山大學
<a href="http://www.ncku.edu.tw/">國立成功大學 / National Cheng Kung University</a>           Oname: 國立成功大學
<a href="http://www.ccu.edu.tw/">國立中正大學 / National Chung Cheng University</a>           Oname: 國立中正大學
<a href="http://www.ym.edu.tw/">國立陽明大學 / National Yang Ming University</a>              Oname: 國立陽明大學
<a href="http://www.ndhu.edu.tw/">國立東華大學 / National Dong Hwa University</a>             Oname: 國立東華大學
<a href="http://www.ntou.edu.tw/">國立臺灣海洋大學 / National Taiwan Ocean University</a>     Oname: 國立臺灣海洋大學
<a href="http://www.ncnu.edu.tw/">國立暨南國際大學 / National Chi-Nan University</a>          Oname: 國立暨南國際大學
<a href="http://sun5.cpu.edu.tw/">中央警察大學 / Central Police University</a>                Oname: 警察大學
<a href="http://www.ntptc.edu.tw/">國立台北師範學院 / National Taipei Teachers College</a>    Oname: 國立台北師範學院
<a href="http://www.tmtc.edu.tw/">台北市立師範學院 / Taipei Municipal Teachers College</a>    Oname: 台北市立師範學院
<a href="http://www.nia.edu.tw/">國立藝術學院 / National Institute of the Arts</a>            Oname: 國立藝術學院
<a href="http://www.ntcn.edu.tw/">國立台北護理醫學院 / National Taipei College of Nursing</a>  Oname: 國立台北護理醫學院
<a href="http://www.ntit.edu.tw/">國立台灣工業技術學院 / National Taiwan Institute of Technology</a>    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
<a href="http://www.tccm.edu.tw/">慈濟醫學院 / Tzu Chi College of Medicine</a>                Oname: 慈濟醫學院
<a href="http://www.cyit.edu.tw/">朝陽技術學院 / Chaoyang Institute of Technology</a>         Oname: 朝陽技術學院
<a href="http://www.yzit.edu.tw/">元智工學院 / Yuan-Ze Institute of Technology</a>            Oname: 元智工學院
<a href="http://www.kpi.edu.tw/">高雄工學院 / Kaohsiung Polytechnic Institute</a>             Oname: 高雄工學院
<a href="http://www.chpi.edu.tw/">中華工學院 / Chung-Hua Polytechnic Institute</a>            Oname: 中華工學院
<a href="http://www.dyit.edu.tw/">大葉工學院 / Da-Yeh Institute of Technology</a>             Oname: 大葉工學院
<a href="http://www.ntcic.edu.tw/">國立臺北商業專科學校 / National Taipei College of Business</a>
                                                                                              Oname: 國立臺北商業專科學校
<a href="http://www.ntcic.edu.tw/">國立臺中商業專科學校 / National Taichung Institute of Commerce</a>
                                                                                              Oname: 國立臺中商業專科學校
<a href="http://www.nptic.edu.tw/">國立屏東商業專科學校 / National Pingtung Institute of Commerce</a>
                                                                                              Oname: 國立屏東商業專科學校
<a href="http://www.ncia.edu.tw/">國立嘉義農業專科學校 / National Chia-Yi Institute of Agriculture</a>
                                                                                              Oname: 國立嘉義農業專科學校
<a href="http://www.niiat.edu.tw/">國立宜蘭農工專科學校 / National Ilan Institute of Agriculture and Technology</a>
                                                                                              Oname: 宜蘭農工專科學校
<a href="http://www.nkit.edu.tw/">國立高雄工商專科學校 / National Kaohsiung Institute of Technology</a>
                                                                                              Oname: 國立高雄工商專科學校
<a href="http://www.ncit.edu.tw/">國立勤益工商專科學校 / National Chinyi Institute of Technology</a>
                                                                                              Oname: 國立勤益工商專科學校
<a href="http://www.lctc.edu.tw/">國立聯合工商專科學校 / National Lien-Ho College of Technology and Commerce</a>
                                                                                              Oname: 國立聯合工商專科學校
<a href="http://www.nypi.edu.tw/">國立雲林工業專科學校 / National Yunlin Polytechnic Institute</a>
                                                                                              Oname: 國立雲林工業專科學校
<a href="http://www.nkhc.edu.tw/">國立高雄餐旅管理專科學校 / National Kaohsiung Hospitality College</a>
                                                                                      Oname: 國立高雄餐旅管理專科學校
<a href="http://ntcpe.ntcpe.edu.tw/">國立台灣體育專科學校 / National Taiwan College of Physical Education</a>
                                                                                              Oname: 國立台灣體育專科學校
<a href="http://www.ntcic.edu.tw/">臺南家政專科學校 / Tainan College of Home Economics</a>     Oname: 臺南家政專科學校
<a href="http://www.tccn.edu.tw/">佛教慈濟護理專科學校 / Buddhist Tz'u Chi Junior College of Nursing</a>
                                                                                              Oname: 慈濟護理專科學校
<a href="http://www.chs.edu.tw/">健行工商專校 / Chien Hsien Institute of Technology and Commerce</a>        Oname: 健行工商
<a href="http://www.vit.edu.tw/">萬能工商專科學校 / VanNung Institute of Technology</a>        Oname: 萬能工商專科學校
<a href="http://203.68.40.3/">南亞工商專科學校 / Nanya Junior College</a>                      Oname: 南亞工商專科學校
<a href="http://gopher.lhjc.edu.tw/">龍華工商專科學校 / Lunghwa Junior College of Technology and Commerce</a>
                                                                                              Oname: 龍華工商專科學校
```

<a href="http://www.mhit.edu.tw/">明新工商專校 / Ming Hsin Institute of Technology</a>	Oname: 明新工商
<a href="http://www.thctc.edu.tw/">大華工商專科學校 / Ta Hua College of Technology and Commerce</a>	Oname: 大華工商專科學校
<a href="http://www.chinmin.edu.tw/">親民工商專科學校 / Chin Min College of Technology and Commerce</a>	Oname: 親民工商專科學校
<a href="http://www.stjctc.edu.tw/">樹德工商專科學校 / Shu Teh Junior College of Technology</a>	Oname: 樹德工商專科學校
<a href="http://www.ccjc.edu.tw/">中州工商專校 / Chung Chou Junior College of Technology and Commerce</a>	Oname: 中州工商專校
<a href="http://203.64.144.1/">建國工商專科學校 / Chienkuo Junior College of Technology</a>	Oname: 建國工商專科學校
<a href="http://www.wfc.edu.tw/">吳鳳工商專科學校 / Wu-Feng Junior College of Technology and Commerce</a>	Oname: 吳鳳工商專科學校
<a href="http://www.ntc.edu.tw/">南台工商專科學校 / Nan Tai College of Technology and Commerce</a>	Oname: 南台工商專科學校

[Organization-Club (Oname)]

<a href="http://140.113.11.235/~gmusic/">台大佳韻音樂社</a>	Oname: 佳韻音樂社
<a href="http://cc.ntu.edu.tw/~b4101009/piano/">台大鋼琴社</a>	Oname: 鋼琴社
<a href="http://med.mc.ntu.edu.tw/~b0401087/chorus/">杏林合唱團</a>	Oname: 杏林合唱團
<a href="http://med.mc.ntu.edu.tw/~b3401006/sinlin/index.htm">杏林弦樂團</a>	Oname: 弦樂團
<a href="http://king.cc.ntu.edu.tw/~b1207031/">基克工作室</a>	Oname: 基克工作室

[Organization-Goverment (Oname)]

<a href="http://expo96.org.tw/">網路博覽會 中華民國館 / Pavilion of Taiwan, R.O.C.</a>	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
<a href="http://www.nccu.edu.tw/zoo/htm/zoomain.htm">台北市立動物園 </a>	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)]

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

Appendix C. Some Mapping Results in 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 Web pages.

[Some Extracted Data in Content Sets before Mapping]

Oname: 資訊新館 63
E-Mail: root@csman.csie.ntu.edu.tw 59

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: 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: 張麗東 155
E-Mail: gdchang@ccms.ntu.edu.tw 160

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	Oname: 資訊新館
E-Mail: focus@www.ntu.edu.tw	Oname: 焦點新聞
E-Mail: news@www.ntu.edu.tw	Oname: 網路新聞
E-Mail: campus@www.ntu.edu.tw	Oname: 校園新聞
E-Mail: tsaibw@ccms.ntu.edu.tw	Pname: 蔡博文
E-Mail: tschiu@ccms.ntu.edu.tw	Pname: 丘台生
E-Mail: ingchen@chem60.ch.ntu.edu.tw	Pname: 陳膺州
E-Mail: yucs@ccms.ntu.edu.tw	Pname: 游張松
E-Mail: hlee@cc.ntu.edu.tw	Pname: 李賢輝
E-Mail: popo@ccms.ntu.edu.tw	Pname: 曾珀雯
E-Mail: kevins@ccms.ntu.edu.tw	Pname: 徐信權
http: http://www.ntu.edu.tw/forest/R17.html	Oname: 國立臺灣大學森林學系暨研究所