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