Proper Name Extraction from Web Pages for Finding People in 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. 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.