自然語言處理技術於中文史學文獻分析之初步應用

劉昭麟*、金觀濤**、劉青峰***、邱偉雲****、姚育松*****

摘　要

自然語言處理是計算機科學中具有相當歷史的學科，過去主要應用於分析與處理現代文字語料。文字作為人類溝通與記錄的主要工具，詞意與語法都與時俱進。因此，處理現代文字語料的計算技術，不見得可以立即應用於歷史語料的處理工作。

本文以中國近現代思想及文學史數據庫為例，實驗如何利用自然語言處理技術輔助史學研究。我們利用 PAT Tree 技術從大量史料中，透過專家的協助來擷取與史學研究相關的詞彙，進一步分析詞彙的語境與共現的現象，最終估計個別文件與研究議題相關度，希望藉此輔助學者以比較有效率的方式，覓得相關的史學文件和分析文件內容。

本文報告兩個應用的初步經驗。我們分析了西元 1905 到 1911 年間的《清末籌備立憲檔案史料》，這一份資料是晚清朝廷討論立憲議題的相關文書；我們也探索了西元 1875 到 1911 年間的《清季外交史料》，這一份文件所牽涉的議題較廣，而我們暫時只研究關於華工的相關史料。這兩項工作關於史學面向的成果提報於本次研討會之另外兩篇論文。

自然語言處理技術固然不能完全取代史學研究者從事史學研究，但是初步經驗顯示，自然語言處理技術有足夠的潛力為史學研究者提供初步分析的服務，讓史學研究者可以比較有效率的方式處理大量的語料，並且把珍貴的研究時間用於知識層次的分析工作。

關鍵字：詞頻分析、共現詞組、齊夫定律、詞彙權重、詞組權重

* 國立政治大學資訊科學系教授，E-mail: chaolin@nccu.edu.tw。
** 國立政治大學講座教授，E-mail: guantao@nccu.edu.tw。
*** 香港中文大學中國文化研究所名譽研究員，《二十一世紀》創刊編輯。E-mail: qingfeng@cuhk.edu.tw
**** 國立政治大學中國文學系博士班研究生，E-mail: acwu0523@gmail.com。
***** 國立政治大學歷史學系碩士班研究生，E-mail: werthersoong@gmail.com。

Analysis of Historical Chinese Documents Using Natural Language Processing Techniques

Chao-lin Liu[*]、Guan-tao Jin[**]、Qing-feng Liu[***]、

Wei-yun Chiu[****]、Yih-soong Yu[*****]

Abstract

Natural language processing (NLP) is a well-known research area in computer science and it has been successfully applied to handle and analyze modern text material in the past. Whether the applications of current NLP techniques can be extended to historical Chinese text is a challenge. Word senses and grammar change over time when people of different times assigned different meanings to the same symbols and word patterns and when different word patterns are used.

The applications of NLP techniques to support the study of historical research based on the text material available at the Database for the Study of Modern Chinese Thoughts and Literature were explored. In recent studies, the PAT Tree method was applied to extract useful Chinese words from the corpora with the help of historians to finalize the keyword selection. The occurrences and collocations of the keywords over the years of interest were also to find a way to rank the historical documents. Hence, historians can find key documents and identify the key sentences more effectively.

The present paper reports the use of NLP techniques to support 2 historical studies. The first discusses how the Qing government attempted to convert itself from a monarchy to a constitutional monarchy between 1905 and 1911 using the documents recorded in *清末籌備立憲檔案史料*. The second issue is the attitude of the Qing government towards overseas Chinese workers during the late 19th century and the early 20th century using the documents recorded in *清季外交史料*. Details about these historical studies are reported in 2 other papers in this conference.

[*] Professor, Department of Computer Science, National Chengchi University. E-mail: chaolin@nccu.edu.tw
[**] Chair Professor, National Chengchi University. E-mail: guantao@nccu.edu.tw
[***] Honorary Research Fellow, Institute of Chinese Studies, Chinese University of Hong Kong; Founding Editor of *the Twenty-First Century* in Hong Kong. E-mail: qingfeng@cuhk.edu.tw
[****] Ph.D. student, Department of Chinese Literature, National Chengchi University. E-mail: acwu0523@gmail.com
[*****] Graduate student, Department of History, National Chengchi University. E-mail: werthersoong@gmail.com.

NLP techniques are not expected to replace the major role of historians in historical studies; however, these techniques should be able to work with historians to improve the efficiency and effectiveness of studies. The preliminary results reported in the present paper and other papers in this conference have suggested the potential of NLP techniques. With the help of computing technologies, historians can delegate some of the searches to computers and spend more time on higher-level thinking than before.

Keywords: frequency analysis, collocation, Zipf's law, term weights, phrase weights