《人文及社會科學集刊》
第二十七卷第四期（104/12）, pp. 683-718

社會科學研究中的文字探勘應用：以文意爲基礎的文件分類及其問題

陳世榮
中國文化大學行政管理學系副教授

隨著電子典藏技術的精進，文字探勘技術逐漸受到重視，本文以社會科學研究在文意區別上的需求，評估監督式機器學習對非結構、複雜文本的分類效果，並就所見問題提出分析與建議。本文從文字探勘與內容分析文意區別上的差異與共通性出發，繼而以新聞報導爲分析資料，針就特定文件意向，遵循一般文字探勘程序，以支持向量機與簡易貝式分類器執行文件分類評估。分析結果指出，文字探勘對於複雜文意的判讀效果值得肯定，但經由共詞網絡分析也發現，文件的編撰風格將影響文件分類的效果。建議研究者在資料處理初期，應反覆評估研究目的、資料特性與分類器模型間的契合度。

關鍵字：文字探勘、文意區別、文件分類、機器學習、共詞網絡分析

Text Mining for Social Studies: Meaning-based Document Classification and Its Problems

Roger S. Chen
Associate Professor
Department of Public Administration and Management,
Chinese Culture University

ABSTRACT

Along with the growing development of electronic information storage, text mining has increasingly gained attention from scholars and practitioners across various disciplines. In response to the need for meaning differentiation in social studies, the study aims to evaluate supervised machine learning classifiers in terms of the performance of document classification. Setting out from the comparison between traditional content analysis and text mining, the evaluation follows a normal procedure of text mining and applies Support Vector Machine and Naïve Bayes classifiers on non-structural, complex social texts extracted from news media. The outcomes of the analysis validate that text mining manages classification well for documents with complex meaning. However, a further co-word network analysis in the study finds that the editing style of data may affect classifiers' performance. It is suggested that, in the early stage of data processing, greater care must be given to the fit between research problems, editing styles, and classifiers.

Key Words: text mining, meaning differentiation, document classification, machine learning, co-word network analysis