大數據分析之文字探勘

以天龍八部為例

Text mining of big data analysis

Take Tianlong Ba Bu as an example

[1] 蔡桂宏　　[2] 顏守玄

[1]Tsai, Gwei-Hung　[2] Yan, Shou-Xuan

[1 2]銘傳大學應用統計與資料科學學系

[1 2]Department of Applied Statistics and Information Science,
Ming Chuan University

摘要

隨著電子典藏技術的精進，文字探勘技術逐漸受到青睞，本研究將文字探勘技術應用在金庸小說上，並以天龍八部為例，利用文字探勘與內容探討對於主要人物出現次數加以各項相關的統計分析。

我們首先收集天龍八部主要人物的姓名、別名並使用 R 語言做為分析工具撰寫程式將所有章回的文字檔逐一匯入，再利用 R 矩陣跟迴圈進行每一個章回主要人物的姓名檢索，最後彙整為一個人物出現次數矩陣，以便進行各項的相關統計分析，包含了長條圖(bar chart)、文字雲(word cloud)、相關係數等比較分析。

關鍵詞：文字探勘、**R** 語言、文字雲、相關分析、天龍八部

Abstract

With the advancement of electronic collection technology, the text mining is gradually favored. This study uses the text mining to apply to Jin Yong's novels, and takes the Tianlong Ba Bu as an example. We use text mining and content dialogue to explore the various related statistical analysis of the number of occurrences of the main characters.

We first collect the names and aliases of the main characters of the Tianlong Ba Bu and use the R language as an analysis tool to read the text files of all the chapters one by one, and then use the R matrix and the loop to carry out the name search of the main characters in each chapter. Finally, it is aggregated into a matrix of character occurrences, in order to carry out various relevant statistical analysis, including comparative analysis of bar chart, word cloud, and correlation analysis.

Keywords: text mining, R language, word cloud, related analysis, Tianlong Ba Bu