利用文本採礦探討《紅樓夢》的後四十回作者爭議

杜協昌[*]

摘要

《紅樓夢》全書共一百二十回。一般公認前八十回的作者是曹雪芹，但後四十回的作者則存有爭議。從前學者們主要是透過可以考定作者、時代、版本的材料，或者從內容前後的連貫性，來推斷後四十回是否為他人所續。隨著電腦的出現，研究者開始利用量化的統計學方法分析前八十回與後四十回之間，是否在用字遣詞上存有顯著的差異。

這類統計方法，通常需先由研究者選定量化標的物（例如虛字頻率），然後再對這些標的物的分佈進行統計檢定。有別於這樣的步驟，本論文運用文本採礦的技術，先讓電腦計算出可能有趣的候選字詞，然後再利用前後綴詞工具來觀察這些字詞的前後，經常相隨有哪些字。我們找到許多前人沒有注意到的字詞，它們在前八十回與後四十回的使用頻率上存在明顯的差異。例如前八十回中有 34 回可看到「嬤嬤」一詞，但「嬤嬤」在後四十回卻一次也沒有出現過。此外，分析「豈」的後綴字，我們發現該字有將近七成是被使用於「豈不」與「豈知」。有趣的是，「豈不」在前八十回的出現頻率明顯偏高，而「豈知」卻僅在後四十回出現。

我們認為，利用資訊工具可以有效幫助人文學者從文本中發掘新事證。我們的實驗結果，支持《紅樓夢》後四十回作者並非曹雪芹的論點。

關鍵字：紅樓夢、作者爭議、文本採礦、詞頻分析、前後綴詞工具

[*] 臺灣大學資訊工程系博士後研究員，Email：tu@turing.csie.ntu.edu.tw。

Using a Text Mining Approach to Study the Authorship Controversy on the Last 40 Chapters of *The Dream of the Red Chamber*

Hsieh-chang Tu[*]

Abstract

The Dream of the Red Chamber is a famous Chinese classic novel written in the 18th century. It has 120 chapters. The author of the first 80 chapters is known to be Cao Xueqin (曹雪芹) but there is a debate on the author of the last 40 chapters. Due to the lack of historical documentation and the rapid development of computer technology, recent researchers turn to use quantitative statistical analysis to solve this problem.

These statistical methods often require the researchers to first choose certain linguistic features, usually function words for Chinese text, and next apply hypothesis testing to check whether the feature frequencies in the first 80 chapters are significantly different from those in the last 40 ones. In this paper we adopt a text mining approach instead. We first define a mining function to spot terms likely to be interesting, and next check these candidates with a prefix-suffix tool to get the actually juicy ones. We find many fascinating terms not noticed by earlier researchers (it is simply too difficult to find these terms without a computer). For instance, the term MaMa (嬷嬷) can be found in 34 among the first 80 chapters, but it never occurs in the last 40 ones. As another example, we analyze the suffix word of unigram Qi (豈) and find that nearly 70% occurrences of this Chinese character are used in one of the two forms QiBu (豈不) and QiZhi (豈知). Interestingly, QiBu (豈不) occurs much more frequently in the first 80 chapters, but QiZhi (豈知) can only be seen in the last 40 ones.

Our experiments show that making use of computer technology can help

[*] Postdoctoral Researcher, Department of Computer Science and Information Engineering, National Taiwan University. Email: tu@turing.csie.ntu.edu.tw.

humanists find interesting facts and clues from text. The results support the statement that the last 40 chapters of *The Dream of the Red Chambe*r were not written by Cao Xueqin (曹雪芹).

Keywords: The Dream of the Red Chamber, Authorship Controversy, Text Mining, Term Frequency Analysis, Prefix-suffix Tool