airiti

東亞觀念史集刊　第六期
2014 年 6 月　頁 327-366

統計偏離值分析於人文研究上的應用

——以《新青年》為例

金觀濤、梁穎誼、姚育松、劉昭麟*

摘　　要

數位方法應用於人文研究，在近幾年逐漸變得重要。數位方法不但增進了學者的研究效率，也提供了另一個角度去思考各種文本。然而，目前仍然有許多技術或方法有待深入探討。本文研究動機運用了齊夫定律偏離值計算來找出文本的關鍵詞，並以《新青年》爲文本來分析是否可行。本研究特點是將關鍵詞的概念加以量化，並從統計理論出發得出一個較客觀的指標。實驗結果發現，利用此統計與數位方法，除了解決主觀判斷的爭議外及減少人工篩選的負擔外，也具有三項突破。一、能夠更加準確地將關鍵詞篩選出來，從而更加細緻地觀察思想變化，以及反映在文本上的語言現象，前者透過「國家」與「青年」這兩個關鍵詞來說明，後者透過「他們」、「我們」、「社會」、「主義」、「階級」、「產階」來說明；二，印證了偏離值大小能夠代表詞的關鍵性，從而揭示出《新青年》的思想變化，從無共識觀念，到

* 作者金觀濤現任「中國近現代思想及文學史專業數據庫（1830-1930）」計畫共同主持人、中國美術學院南山講座教授；梁穎誼現爲政治大學統計學系博士候選人；姚育松現爲廣州中山大學哲學系博士候選人；劉昭麟現任政治大學資訊科學系特聘教授。

思想討論，最後到意識形態確立的過程；三，印證了關鍵詞詞數能夠部分反映文本的思想變化，由於觀念濃縮於關鍵詞，關鍵詞的多寡能夠反映明顯觀念的多寡。

關鍵詞：齊夫定律、偏離值、《新青年》、關鍵詞

airiti

airiti

Application of Statistical Residual Analysis to Humanities Studies:

Using *Xin Qing Nian* as Example

Jin Guan-tao, Leong Yin-yee,

Yu Yih-soong, Liu Chao-lin

Abstract

Application of digital methods on art and humanity studies had become crucial research in recent years. Digital methods can help to improve the efficiency of studying, and also to have another point of view to rethink the materials. However, We have a lot of issues that need to be studied in details. Our aim is to proposea residuals calculation based on Zipf's law, in order to find out the key-words of the materials with using *Xin Qing Nian* as an experiment example. We redefine the concept of key-words in a quantitative way, and establishing anless-subjectivecriterion from statistical theory. The experiment had found out that, as using statistic and digital method, we have three breakthroughs others from diminishing the criticality controversy by subjective judgments and the burdens of manual Selection. First, as the reason of be able to assure

airiti

more accurately of the key-words, we can have more precise observation to thoughts changing and the language phenomena. Second, giving an evidence of how the importance of the key-words can be demonstrated by residual quantities. Third, giving an evidence of how the quantity of key-words can more or less demonstrate the emerging of obvious ideas.

Keywords: Zipf's Law, Residual, *Xin Qing Nian*, Key-words

airiti

統計偏離值分析於人文研究上的應用

——以《新青年》為例*

金觀濤、梁穎誼、姚育松、劉昭麟

一、前言

數位人文的研究目的，是要藉助數位方法推進人文研究的拓展已是學界共識，早在去年，臺灣政治大學「歷史與思想數位人文實驗室」研究團隊就已經利用齊夫定律來確定文本關鍵詞的研究，本文即是站在此成果的基礎上，結合統計、資科、人文三項專業，進一步透過齊夫定律（Zipf's Law）的偏離值計算，來確定文本的關鍵詞，從而展開人文解釋。

應用齊夫定律來確定關鍵詞的假定爲如此。齊夫定律作爲一經驗法則，發現普遍上文本的詞頻（frequency）之取對數（log）之後，便會有一特定的分布，因此能夠在知道某文本的字數之前提下，在還未實際統計該文本的詞頻前，便能預設該文本應有的理論曲線，即爲

* 本文研究中關於《新青年》的詞彙資料，取自「中國近現代思想史專業數據庫（1830-1930）」（香港中文大學中國文化研究所當代中國文化研究中心開發，劉青峰主編），現由臺灣政治大學「中國近現代思想及文學史專業資料庫（1830-1930）」計畫（鄭文惠主編）持續開發功能與完善數據庫並提供檢索服務，謹致謝忱。此外，文中有關統計的部分，亦頗受益於政治大學統計學系余清祥教授的指導，特此表達謝意。

airiti

其按照齊夫定律應有的分布。想當然的是，實際分布曲線與理論曲線必定是有差異的，而此差異即表現在某些樣本，也就是某些詞之詞頻偏離了理論曲線，這通常表現在某詞之詞頻被異常地大量或少量使用。我們將那些詞頻遠遠超出其他詞彙使用量的詞彙，假定爲該文本的「關鍵詞」。可以發現到，詞頻異常高的樣本群，其實際曲線與理論曲線通常呈現出一明顯的偏離段，由此我們便假定這些樣本群即爲「關鍵詞詞叢」。

過去我們基本或者說初步證實了以上的假定是可以成立的。首先在技術上，由於中文不似英文，不能直接斷詞，因此我們利用了 PAT-Tree 技術（McCreight，1976），假定重複出現二次以上的連續字串即爲一詞，由此提取了文本所有可能詞彙的詞頻，製作出文本詞頻的實際對數圖，從「準詞彙表」中，透過人工篩選來確定「關鍵詞詞叢」以展開人文研究。[1]我們分別以《清季外交史料》及《清末籌備立憲檔案史料》爲文本，前者透過共現關鍵詞的分布，發現到近現代中國的華人觀念的形成，在相當大的程度上，與保護華工的外交糾紛有關；[2]後者則透過關鍵詞來找出關鍵文本，解釋清末籌備立憲失敗的背景，與行政機關不斷遭受諮議局的挑戰有關。[3]

然而，過去的研究屬於初步的探索，有兩大未盡善處，其一是樣本偏離段並非在對照理論曲線的情況下得出，故此只憑詞頻高低來判斷，對於何謂關鍵詞並沒有完全根據樣本分布來區別，可說無可避免地相當主觀；其二是基於只憑人工判斷的關係，爲了盡量囊括可能的

---

1 參見劉昭麟、金觀濤、劉青峰、邱偉雲、姚育松：〈自然語言處理技術於中文史學文獻分析之初步應用〉，《第三屆數位典藏與數位人文國際研討會論文集》（臺北：臺灣大學數位典藏研究發展中心，2011 年）。

2 參見金觀濤、邱偉雲、劉昭麟：〈「共現」詞頻分析及其運用——以「華人」觀念起源爲例〉，《第三屆數位典藏與數位人文國際研討會論文集》（臺北：臺灣大學數位典藏研究發展中心，2011 年）。

3 參見金觀濤、姚育松、劉昭麟：〈社會行動的數位人文研究：以清末預備立憲爲例〉，《第三屆數位典藏與數位人文國際研討會論文集》（臺北：臺灣大學數位典藏研究發展中心，2011 年）。

關鍵詞，所判斷的偏離詞段往往過長，以致於準關鍵詞詞表的詞數太多，動輒上千上百，增加了研究者篩選無意義字串的負擔。本文即是抱著進一步驗證齊夫定律的使用可能性，以及解決上述兩大問題的研究動機上展開研究。

要解決上述兩大問題，就需要能夠精確判斷偏離段的方法，也就是計算「偏離值」。所謂「偏離值」，即是在座標上，實際曲線與理論曲線的差距值。某樣本的偏離值越大，即代表該樣本所表示的詞彙，越有可能是關鍵詞。我們以《新青年》爲實驗對象，發現到這樣的對照，可說是在數位與人文研究的結合上，取得了巨大的進步，解決了上述的兩大問題。第一，透過偏離值的計算，我們能夠確定文本的最大偏離段爲何，避免人工判斷的主觀問題；第二，由於能夠確定最大偏離段，便不需要抱持著盡量囊括所有可能詞彙的擔心，過去動輒上百的關鍵詞詞數的現象已經不再出現，本文分別對《新青年》十一卷每卷分析，發現最多的關鍵詞詞數的第7卷只有30個，而最少的第1卷只有2個，大大減輕了研究者篩選無意義字串的負擔。

此外，實驗結果也發現，所找出的十一卷的每卷關鍵詞詞表，比較對照下，相當明顯地反映了《新青年》的思想變化，與過去的人文研究結果可謂相當一致，由此進一步證實了齊夫定律能夠適用於分析文本觀念結構，並且還能夠更加精確地捕捉關鍵詞，把看似不像反映概念或觀念，而更像常用詞的關鍵詞，例如「我們」、「他們」，在各個關鍵詞消長變化的綜合對照下，發現這些經常被忽略的常用詞其實也能解釋某種觀念的變遷。

更加令人驚喜的是，除了解決之前研究不足之外，加入了精確偏離值計算的結果，也帶來了兩大突破。第一，樣本偏離值越大即爲關鍵性越大，我們從《新青年》發現，偏離值隨著卷期遞增而逐漸上升，例如從第1卷的最大偏離值是15.01，到第6卷的28.12，最後到第11卷的64.17，關鍵詞是濃縮概念或觀念的載體，這種關鍵性的逐漸增加，也就表示了觀念形塑的確立過程；第二，關鍵詞詞數也

airiti

能反映時代特性，從《新青年》發現，詞數是從少到多，又從多到少，可以這麼解釋，關鍵詞詞數少，即表示當時人們共有的代表觀念少，關鍵詞詞數多，即表示當時人們共有的代表觀念多。而關鍵詞詞數在《新青年》的變化分布，正是證實了這一解釋，從第1卷到第6卷（1915年9月到1919年11月），關鍵詞詞數平均爲9個，第7卷（1919年12月到1920年5月），關鍵詞詞數爲30個，第8卷到第11卷（1920年9月到1926年7月），關鍵詞詞數爲14個，這樣的樣本現象正好解釋了當時中國人從難以找到新思想的貧乏中（故此前期關鍵詞詞數少），經過社會主義引入而產生的思想激盪中（故此1919年到1920年的關鍵詞詞數多），最後確立意識形態（故此後期關鍵詞詞數又變少）。

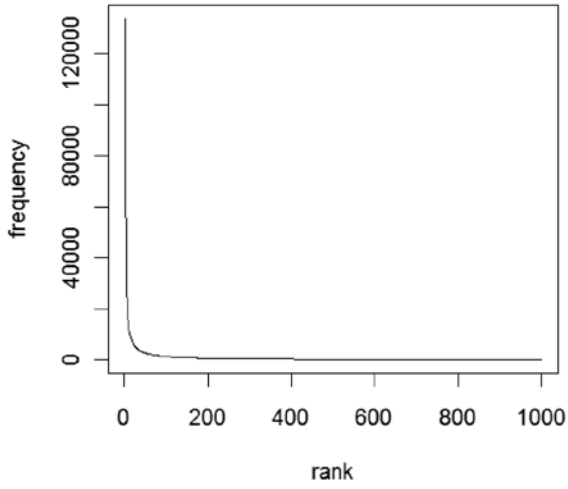
這兩大突破可以說是將數位人文研究提高到新的高度，已經不只局限於從解釋關鍵詞的意義來解釋文本、發現歷史，而是在還未進行文本解釋前，便能夠透過觀察關鍵詞的偏離值及詞數，來擁有一個「文本所可能反映時代特徵之線索」（必須強調是「可能」及「線索」，因有待於研究更多文本來驗證），可說爲人文研究提供大大的方便。

以下，本文將透過三節來說明這次的突破，首先第一節是方法論，解釋理論曲線、偏離值的計算及相關調整；第二節是資料分析，將呈現出《新青年》每卷的關鍵詞詞表；第三節是文本分析，將說明所找出的關鍵詞詞表如何有效地反映了與過去人文研究的一致及突破。

二、統計模型與研究方法

在本研究，最主要的想法爲運用實際資料與齊夫定律的差異，進而找出偏離最大段的部分。並且，我們認爲關鍵詞會異常的多量出現。基於這個想法，我們從統計理論上，建立出一套分析方法。在本節，爲了閱讀方便，我們不列出繁瑣的公式，只簡明扼要闡述此理論的概念與邏輯。其背後的數學公式將會列在附錄四。

齊夫定律自 1935 年由 George Kingsley Zipf（1902-1950）提出來後，在語言學，社會學都有廣泛的應用。[4] 該定律被廣泛運用的原因在於，許多文學現象、社會現象、自然現象都出現了類似的規律。齊夫定律主要在敘述某組物件經過頻次排序後，頻次開始會快速的遞減，但是到了後面遞減速度會變的越來越慢。舉例來說，若將全國的城市人口進行排序再把圖畫出，將會發現排序的分布接近圖一的曲線。這其中有兩個主要現象：一、大都市的數量很少，但小城市的數量卻很多；二、前幾大城市的人口差距很大，但是排名後面的城市彼此規模其實都相差不遠。此現象普遍被認爲是一種經驗法則，也就是說儘管造成此現象的原因有不同的解釋，但先前很多資料都發現，齊夫定律都大概能準確的描述這些現象。



圖一：城市規模的齊夫定律曲線
橫軸爲城市排名，縱軸爲人數

[4] Bruce M Hill, “Zipf’s law and prior distributions for the composition of a population”, *Journal of the American Statistical Association*, 65 (1970): 1220-1232; George A Miller, *The Science of Words*. Scientific American Library, a division of HPHLP, distributed by W.H. Freeman and Company, 1991.

airiti

在自然語言的處理上，齊夫定律的運用也有一段長遠的時間。至今相關的研究數量不勝枚舉。[5]這是由於大量的研究發現，語言資料如文本、演講稿也大致符合齊夫定律。[6]其中，對於英文語言的研究相對較多。不過在後期，部分文獻也針對中文語言進行分析。[7]

後來，有學者在齊夫定律的架構上，發展出其他的定律。其中之一爲 Zipf-Mandelbrot 定律。由於 Zipf-Mandelbrot 定律比齊夫更有彈性，因此，後續的研究，也有學者逐漸考慮使用 Zipf-Mandelbrot 定律去描述各種現象。同樣的，我們在中文文本也發現 Zipf-Mandelbrot 能更精確的描述文字分布。圖二描述了《新青年》某一卷的詞彙分布。不難發現，總體上 Zipf-Mandelbrot 的曲線更加的貼近實際的詞頻分布。因此，本研究將使用 Zipf-Mandelbrot 定律。

然而，雖然在宏觀面來說 Zipf-Mandelbrot 法則能夠很好的描述文字現象，但是在微觀上，排序很前面的字卻跟 Zipf-Mandelbrot 法則有一段不小的差距。我們認爲這正符合了強調觀念的語言現象。換句話說，若一個具有中心思想的文本，必會一直強調與其觀念相關的關鍵字，導致這些字會異常的增多，偏離經驗法則該出現的理論次數。這也是本文所提出方法的精神所在。針對此現象，本文將提出這種偏離程度的計算方式（附錄四），稱之爲偏離值。其解讀方式非常

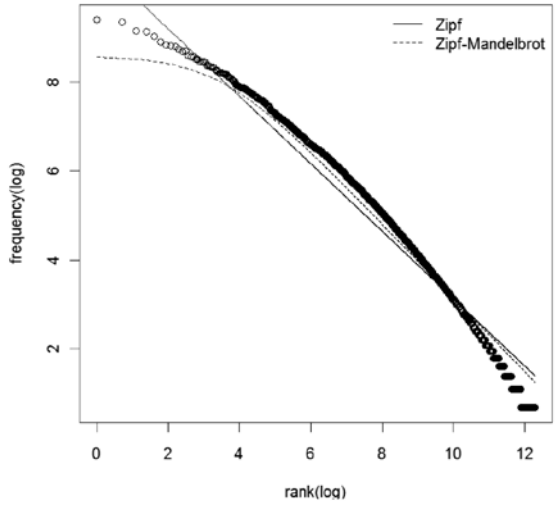
---

5 Colin Martindale, SM Gusein-Zade, Dean McKenzie, and Mark Yu. Borodovsky, “Comparison of equations describing the ranked frequency distributions of graphemes and phonemes”, *Journal of Quantitative Linguistics*, 3(2) (1996):106-112; DR Ridley, EA Gonzales, “Zipf’s law extended to small samples of adult speech”, *Percept. Mot. Skills*, 79 (1994):153-154; S Naranan, VK Balasubrahmanyan, “Models for power law relations in linguistics and information science”, *Journal of Quantitative Linguistics*, 5 (1998):35-61.

6 有研究者將相關的研究整理在網頁上：http://ccl.pku.edu.cn/doubtfire/NLP/Statistical_Approach/Zip_law/references%20on%20zipf's%20law.htm

7 R Rousseau, Qiaoqiao Zhang, “Zipf’s data on the frequency of Chinese words revisited”, *Scientometrics*, 24(2) (1992):201-220; Hang Xiao, “On the Applicability of Zipf’s Law in Chinese Word Frequency Distribution,” *Journal of Chinese Language and Computing*, 18(1) (2008):33-46.

直觀，當某字的偏離值越大，表示該字異常的多，反之則正常。另外，爲了判斷哪些字爲「異常」，我們也引進了統計上常見的盒裝圖（Boxplot），助以設立一個門檻值來篩選「異常」詞彙。具體的操作方式，在下一節會加以說明。



圖二：《新青年》的語言分布
小圈爲實際詞頻

事實上，我們提出的分析方法非常類似統計分析裡的適合度檢定（Goodness-of-fit test），或迴歸模型（Regression model）上的殘差分析（Residual analysis）。其精神也是嘗試分析資料上的反常觀察值。以上所提及的方法與公式，在一般的統計書籍都會提到，這裡就不多闡述。

在本節最後，我們綜合以上的說明，整理成本文方法的標準分析流程：

1. 把需要分析的文本，使用 PAT-Tree 技術進行切詞。並將切詞結果統計成詞頻統計表，如附錄二。

airiti

2. 使用附錄四的公式計算出每一個字的偏離值。

3. 運用盒裝圖，或其他方式判斷哪些字爲「異常」的多。

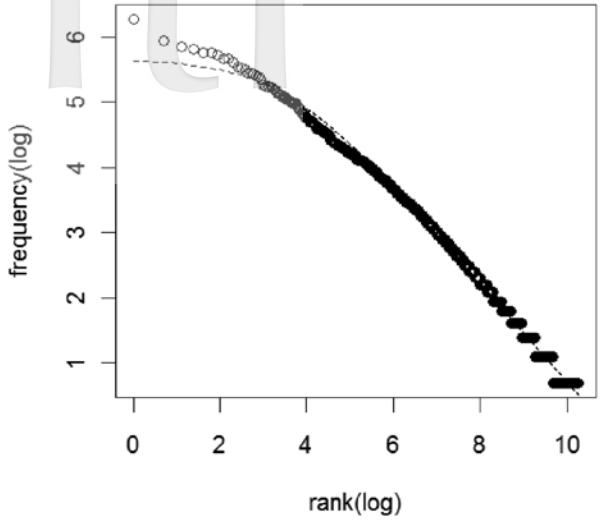
4. 人文學者介入，篩掉常用字、慣用字、並挑除與觀念密切相關的關鍵字，接著進行歷史、人文詮釋。

三、《新青年》數據分析

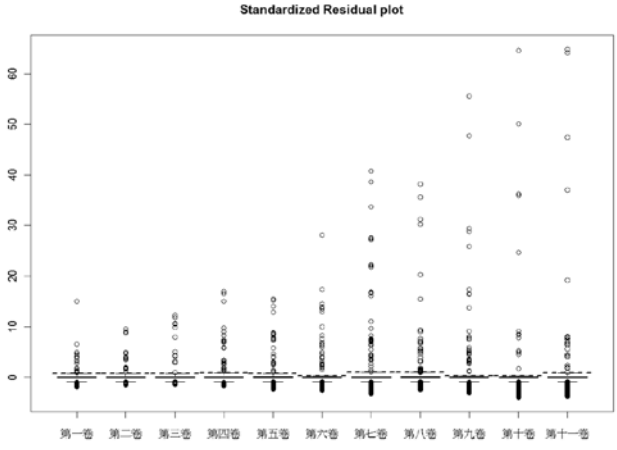
在這一節將呈現《新青年》的分析結果。首先採用 PAT-Tree 技術來提取詞頻，分別以《新青年》十一卷作爲文本，將每卷的所有連續出現 2 次以上的二字字串的詞頻統計出來，並按照詞頻由高至低排列，製作出「準詞表」。之所以只統計出二字字串，是因爲中文詞彙結構複雜，不同字串的詞彙所指涉的概念是非常不一樣的，綜合起來一起分析未必適於初步的實驗。另外，從過去的研究發現，中文以兩個字來表示概念或指涉是最常見的。總而言之，只統計二字字串可便於控制變數。

使用 PAT-Tree 擷取了所有可能的詞彙之後，便針對《新青年》十一卷上的詞（Bi-gram）個別分析。《新青年》在每一卷上的總詞數大約爲二十多萬。由於資料不算少，分析結果也會比較穩定與可信。在實際分析上，我們使用 R 統計軟體內去計算附錄四的公式，包括了上述所提及參數的最佳解。

圖三展示了 Zipf-Mandelbrot 對《新青年》第 1 卷配適狀況。虛線爲理論值，小圈爲實際值。其他卷的 Zipf-Mandelbrot 配適圖與第 1 卷類似（附錄三）。可以發現，配適狀況總體上不算太差。然而，在前段的字數都異常的高，這也是偏離 Zipf-Mandelbrot 最大的部分。我們進一步把偏離值用盒裝圖畫出。圖四爲其結果。



圖三：《新青年》第 1 卷的實際值（小圓圈）與理論值（虛線）



圖四：《新青年》第 11 卷偏離值盒裝圖[8]

[8] 小圓圈：每詞的離群值；虛線：Tukey 建議的門檻值。

airiti

接下來，我們需要判定那些詞是待分析的候選詞，在這裡，我們使用了統計的離群值（Outlier）概念去切入。根據 Grubbs（1969）的定義，離群值可視爲與樣本不一樣的個本。[9]然而，在統計這半個世紀的發展以來，各種判別離群值的方法陸續被提出。其中，常被使用的方法爲 John Wilder Tukrkey (1915-2000)（1977）的方法。Tukey 提出，只要任何一樣本點超過$[Q1-(k\times IQR), Q3+(k\times IQR)]$ 這個區間，[10] 就可視爲可能的離群值，至於 k 值他則建議取爲 1.5 或 3。

圖四表示了偏離值分散的狀況。很明顯的，某些詞的偏離值特別大。這表示這些詞出現的頻率異常的多。我們先使用 Tukey 的建議，取 k 爲 3，找出可能的候選詞，圖四中，橫的虛線表示了 Tukey 的門檻值。我們可以發現，每一卷都有部分的詞被列爲離群值。爲了方便討論其歷史意義，在這些離群值中，我們優先選取離群值特高的詞，從盒裝圖上，可以很明顯地看出各卷大約在 5-10 之間開始，就有一群詞的偏離值特別高。我們將這些詞挑選出來後（附錄一），接著由人工挑除一些常用關鍵字如「今日」、「以爲」等。這類型的字於詮釋觀念並無作用，純粹只是維持句子的完整性。最後再由人文學者挑除與觀念密切相關的關鍵字，由此確立出各卷的關鍵詞詞表。另外，有趣的是，異常大的偏離值在後期頻密出現，這部分在下一章節會詳加探討。

在接下來的章節，我們將從人文研究的角度切入進行探討，而表一則爲下文所關注的關鍵詞。另外，爲了簡便，接下來的 Zipf-Mandelbrot 定律，我們一律簡稱齊夫定律。

---

[9] 原文爲：“An outlying observation, or ‘outlier,’ is one that appears to deviate markedly from other members of the sample in which it occurs.”

[10] Q1、Q3 分別爲第一跟第四四分位數，IQR = Q3 – Q1。

airiti

四、《新青年》數據結果在人文研究上的反映

可以這麼歸納，這次對齊夫定律偏離值計算應用於人文研究上的實驗，可歸結爲三大發現：一、關鍵詞與觀念變遷的關係；二、偏離值與觀念確立的關係；三、關鍵詞詞數與觀念確立的關係。以下分述於三小節：

（一）關鍵詞與觀念變遷的關係

我們知道，《新青年》所反映的思想變化見證了一代中國知識群體從新文化運動到接受社會主義、馬克思主義的一段觀念形塑過程，這次實驗所得出的關鍵詞詞表，也正好反映了這段形塑過程的關鍵詞變化，從關鍵詞表（附錄一）可以發現到代表社會主義、馬克思主義的關鍵詞，如「勞動」、「階級」、「資本」、「共產」、「無產」都在第6卷（1919年1月15日至1919年11月1日）之後才出現。僅止於這一發現，只能說明以齊夫定律尋找關鍵詞是可行的，本文還要提出的是，加入偏離值計算的齊夫定律實驗，單純以數據模型來判斷關鍵詞詞段，從而排除了主觀判斷的誤差，能夠更加精確地捕捉到關鍵詞。

在利用數位技術上，研究者單純以詞頻高低來決定關鍵詞，這是最主觀的作法。儘管從附錄一可看出，偏離值越高者，其詞頻即越高，看似不需要計算偏離值即可憑詞頻來判斷關鍵詞。然而，偏離值是在比較理論曲線與實際曲線所得的結果，單憑詞頻高低，我們無法清楚辨別哪些詞段相對於其他詞彙，才是關鍵的。每個文本都有爲數眾多的詞彙，這種基本作法只能讓研究者明顯看到詞頻排在前面一、二位的關鍵詞，至於從哪一詞頻開始，才能算作關鍵詞詞段，研究者根本很難確定。基本作法中研究者一開始看到的詞表，也就是採用PAT-Tree擷取詞頻而製作出的準詞表，便如附錄二。附錄二是《新青年》第1卷的詞頻排列表，在本文只羅列前面一百個，但事實上原

airiti

始的表是有28,768 個詞。基本作法中，爲了盡可能地囊括所有關鍵詞，因此研究者不得不有這樣的判斷：從詞頻高者往下看詞頻低者，一直看到不再或甚少出現「看起來像關鍵詞」的詞彙後，便判斷該詞彙之詞頻以上者，即爲關鍵詞詞段，但何謂關鍵詞，如以主觀判斷的話，就是公說公有理，婆說婆有理。以附錄二爲例，甲研究者可以宣稱排名35 的「共和」是最後一個「看起來像關鍵詞」的詞彙，但乙研究者也可以宣稱排名71 的「民族」才是最後一個「看起來像關鍵詞」的詞彙，兩者的說法都有其道理，因爲我們知道當時中國人的願望就是要建立「共和」,「民族」自強。

爲了避免主觀問題，我們採用齊夫定律之偏離來進行判斷。齊夫定律假定了文本的詞頻分布在取得對數之後應有一條理論曲線，我們便是根據實際曲線與理論曲線之差異，來確立最大偏離段。從附錄三的圖可看出，幾乎每一卷的曲線在前頭部分，也就是高詞頻的部分，都會明顯看出理論曲線與實際曲線的巨大偏離，進而使用盒裝圖（圖四）加以區分後，便會明顯看出偏離段爲何，也就是最大空隙之後的樣本，從圖四可看出，卷一的最大偏離段爲前面兩個樣本，也就是表一中的「政府」與「國家」。

這裡必須注意，在設定門檻值時，雖然我們使用的是3 倍 IQR 的方法，但是我們可以從統計理論了解一個原則，超過三個標準差或以上的事件，其發生機率是非常低的。所以，門檻不應設比其更低的值。然而，另一方面，我們發現超過三個標準差的此種事件，偶而會出現較多，這會造成人文上歸納解讀的困難，這時，門檻值可依資料的形態稍微調高。

用齊夫定律來確立關鍵詞，不只減少了人工篩選的負擔，也排除了肉眼判斷所可能帶來的爭議，更能夠明確或說精準地掌握到文本特有的關鍵詞，從而更加細緻地觀察到觀念的變遷，本文以下表中的兩個例子來說明。

airiti

表一：《新青年》中，本文所提到的關鍵詞在各卷的出現，黑點表示該字在當卷爲關鍵詞。

卷數	一	二	三	四	五	六	七	八	九	十	十一
國家	●										
政府	●										
青年		●									
娜拉				●							
文學			●	●							
中國			●	●	●		●				
社會			●	●		●	●	●	●	●	●
我們				●	●	●	●	●	●	●	●
他們					●	●	●	●	●		
主義						●		●	●	●	●
資本								●	●	●	●
階級								●	●	●	●
革命									●	●	●
產階									●	●	●

1.「國家」與「青年」

第1卷的關鍵詞是「國家」，第2卷的關鍵詞是「青年」。這樣的結果是有點令人出乎意料的。我們知道，《新青年》的創刊目的即是要教育青年，在創刊號上，作爲主編的陳獨秀（1879-1942）便撰寫了一篇〈敬告青年〉，強調青年精神代表朝氣，因此中國之存亡繫於青年是否能夠自強，由此提出六點意見予青年：

> 惟屬望於新鮮活潑之青年，有以自覺而奮鬥耳。自覺者何？自覺其新鮮活潑之價值與責任，而自視不可卑也。奮鬥者何？奮其智能，力排陳腐朽敗者以去，視之若仇敵，若洪水猛獸，而不可與爲鄰，而不爲其菌毒所傳染也。[11]

[11] 陳獨秀：〈敬告青年〉，《新青年》第1卷第1期，1915年9月15日。

airiti

那麼爲何「青年」一詞要到第2卷才成爲關鍵詞呢？可以發現到，教育青年之重要，在第1卷裡，是放在振興國家的大原則下才被強調的，例如高語罕（1888-1948）便說明瞭這樣的關係，在〈青年與國家之前途〉[12] 中，便說明國家之強大有賴於國民之自強，而能夠期待之國民，也只有青年：「內以刷新政治，鞏固邦基；外以雪恥禦侮，振威鄰國，則捨我青年誰屬？蓋民爲國之根本，而青年又民之中堅也。欲國之強，強吾民其可也；欲民之強。強吾青年其可也。」

那麼我們是否便能夠這樣解釋，第1卷的關鍵詞之所以不是「青年」，而是「國家」與「政府」，是因爲青年之教育是放在國家強大的原則下才被強調，故此對於國家與政府的強調自然更多？當然這樣的解釋是可以成立的，但是仍然解釋不了爲何到第2卷的時候，「國家」與「政府」便不再是關鍵詞。繼續探索下去，會發現到一段微妙的觀念轉折，即歐戰對時人的影響。

《新青年》的創刊目的是教育青年以求國家自強，在初期中，國家如何自強是被簡單歸結到依賴青年自覺，以及學習西方。新文化運動具有全盤反傳統的觀念結構，在此觀念底下，傳統的、老舊的、年老的都被視爲不進步，而西方的、現代的、年輕的都被視爲進步。例如汪叔潛（生卒年不詳）在〈新舊問題〉[13] 便強調，「上自國家，下及社會，無事無物，不呈新舊之二象。吾人與事物之緣，一日未斷，則一日必發生新舊問題」，以及「所謂新者無他，即外來之西洋文化也。所謂舊者無他，即中國固有之文化也。如是，則首當爭辨者，西洋文化與中國文化，根本上是否可以相容」，強調「亦必知歐美各國之家族制度、社會制度，以至於國家制度，固無一焉可與中國之舊說。勉強比附者也」。

這樣簡單的二分法，是因爲時人認爲已有一既定要遵守的普遍原則，即西方建立民族國家之理論，因此在第1卷「國家」與「政

[12] 高語罕：〈青年與國家之前途〉，《新青年》第1卷第5期，1916年1月15日。
[13] 汪叔潛：〈新舊問題〉，《新青年》第1卷第1期，1915年9月15日。

airiti

府」是爲關鍵詞，乃因爲當時《新青年》的作者們正大量地介紹西方建構民族國家的理論，在他們看來，中國的問題就是因爲帶著過去的傳統物，不像一個眞正的國家，而他們的願望就是要按照西方理論建立一個眞正的「國家」，擁有一個眞正的「政府」，「國家」與「政府」兩字蘊含著普遍性質，即當作者期許「中國要如何如何」，與他說明「國家是如何如何」或「政府是如何如何」，兩者的語境是一樣的，因爲中國的問題就是不能按照西方理論成爲「眞正的國家」或「眞正的政府」。

例如高一涵（1885-1968）便說道：「……皆爲吾國道德之格言。今按國家原理與世界潮流，始無一不形其牴觸，功利家欲唱爲廢棄道德說者，蓋亦有不得已之苦衷云」，[14]「國家原理」是與作爲普遍原則的「世界潮流」相一致的。

然而，歐戰的慘烈使人發現到西方國家理論所可能帶來的危害，從而意識到不是只要建立所謂眞正的國家，就能夠臻於幸福，中國最大的危機正是來自各個國家的競爭。劉淑雅（1890-1958）便發出悲鳴：「今而後方知戰鬥乃人生之天職，和平爲癡人之迷夢」，並意識到西方帝國主義的潛藏危機：「此次歐洲戰爭終局以後，即爲黃白人種陣師鞠旅以決生死之時期」。[15] 可以說，歐戰的陰影讓中國人一切追求西方的願望破滅，轉而視西方爲洪水猛獸，由此意識到自身與西方的不同，加上戰後談判中國作爲戰勝國竟無法取回山東半島的權益，更深深刺激了時人的神經，由此西方國家理論是否能夠作爲一普遍原則便得到質疑。在此背景下，「國家」及「政府」所蘊含的普遍性質，便被消解。

學習西方理論的潮流，在此之後，被新文學運動所取代，從第3

---

[14] 高一涵：〈共和國家與青年之自覺〉，《新青年》第1卷第1期，1915年9月15日。

[15] 劉叔雅：〈歐洲戰爭與青年之覺悟〉，《新青年》第2卷第2期，1916年10月1日。

airiti

卷到第5卷（1917年3月1日至1918年12月15日）的關鍵詞便可看出：「中國」、「文學」、「娜拉」等等。值得注意的是，文學運動的成功使白話文取代文言文，於是「我們」、「他們」便被經常使用，而成爲關鍵詞。

2.「他們」、「我們」、「社會」、「主義」、「階級」、「產階」

「我們」與「他們」無庸置疑的是，是作爲一種區別我者與他者的常用詞。可以發現到，「我們」與「他們」隨著時間展延，其偏離值越趨下降，取而代之的是「主義」、「階級」、「社會」，「資本」。這是爲什麼呢？這恰好顯示了意識形態確立過程的現象。

想像一下，人們之所以要區別我者與他者，是要說明我者與他者之不同，例如「我是什麼什麼」，「他是什麼什麼」，其中的「什麼什麼」便是爲區別我者與他者不同的形容詞，本文姑且稱之爲「分別形容詞」，而「我們」與「他們」，本文姑且稱之爲「分別主詞」。例如「我們是好人」，「他們是壞人」，「好與壞」便是分別形容詞。

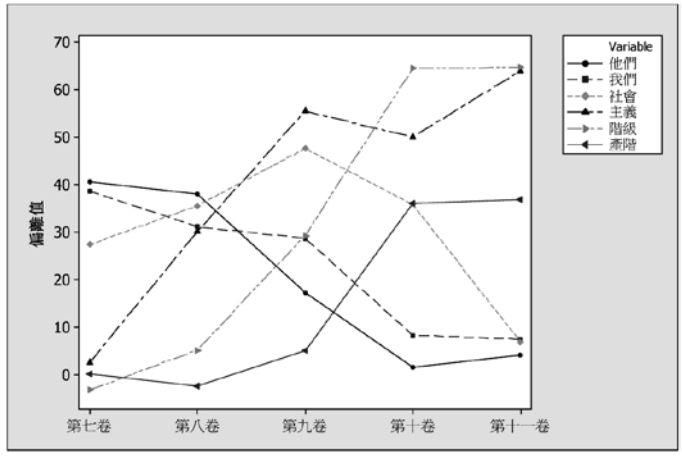
可以發現到，「主義」與「社會」作爲社會主義的意識形態用語，既然是代表意識形態，便意味著透露我者或他者的某種特性，於是我們便可將之定義爲「分別形容詞」，其成爲關鍵詞首先出現於第7卷（1919年12月1日至1920年5月1日），這正是中國引入馬克思主義的發生年，由李大釗（1889-1933）的〈我的馬克思主義觀〉[16]開其端。此外，屬於意識形態用語，但可作爲分別主詞的是「階級」及「產階」，所謂「產階」即出自於「資產階級」、「無產階級」、「有產階級」（因爲本文採用的PAT-Tree只計算二字連用字串，故此排除了四字連用字串）。

從第7卷到第11卷（1919年12月1日至1926年7月25日），可以發現到分別形容詞偏離值逐漸高出分別主詞的趨向，甚至有分別主詞被取代的趨向，這表現在「我們」與「他們」的偏離值逐步下

[16] 李大釗：〈我的馬克思主義觀〉（上下），《新青年》第6卷第5、6期，1919年5月及11月1日。

airiti

降，「主義」與「社會」的偏離值逐步上升，「階級」及「產階」的偏離值也逐步上升。表二列出了這數個關鍵詞的頻次分布。可明顯發現，在這幾卷的前、中、後期的關鍵詞結構比例迥然不同。爲了進一步證實這個現象，我們使用卡方適合度檢定去檢定其齊一性（Test of homogeneity）。[17] 齊一性檢定在統計學上使用甚廣，其目的正是去檢驗不同群體、時期的結構比例是否相同。透過檢定，我們發現這以第9卷區隔的三個時段，用詞頻率確實呈現了統計上的差異。[18] 請見以下圖表（圖五）與（表二）：



圖五：《新青年》第7卷到第11卷的偏離值：「他們」、「我們」、「社會」、「主義」、「階級」、「產階」

[17] 一般的初等統計學課本都會提到齊一性檢定的詳細過程，內容通常會在「卡方分析」的章節內。

[18] 使用卡方檢定，其卡方值爲8490.57，自由度爲10，p-value 則小於0.001。

airiti

表二：《新青年》第7卷到第11卷「他們」、「我們」、「社會」、「主義」、「階級」、「產階」的頻次分布

	第7、8卷	第9卷	第10、11卷
他們	3705	1127	1880
我們	3410	1474	2382
社會	3195	1998	3300
主義	2119	2231	5998
階級	973	1472	6448
產階	272	699	4427

這顯示了意識形態確立過程中會出現的語言現象：一是表現在分別形容詞比起常用分別主詞，變得越加被常使用；一是表現在一旦意識形態確立後，屬於此意識形態下用於區別我者與他者的概念用語，亦即特有的分別主詞，會取代常用分別主詞。從圖五可以發現，這樣的轉折點明顯發生在第9卷（1921年5月1日至1922年7月1日）。眾所周知的是，中國共產黨正是建立於1921年，從此《新青年》成爲中共的機關報，主編陳獨秀便於第9卷第3期發表文章〈隨感錄：一二四，政治改造與政黨改造〉，[19]表明共產黨對改造中國的重要，以及顯示了「階級」作爲意識形態分別主詞的例子：

> 有人說，在有產階級的政治之下，由金力造成的政黨，這種現象是必然的，是無法改造的，只有以共產黨代替政黨，才有改造政治底希望。我以爲共產黨底基礎建築在無產階級上面，在理論上，自然要好過基礎建築在有產階級上面用金力造成的政黨；……。

由此可見，中國共產黨的成立是無產階級革命意識形態確立的象徵，印證於關鍵詞詞表的變化，可發現完全符合。

[19] 陳獨秀：〈隨感錄：一二四，政治改造與政黨改造〉，《新青年》第9卷第3期，1921年7月1日。

airiti

（二）偏離值與觀念確立的關係

在前言中已經說明偏離值即爲樣本的實際曲線與理論曲線在座標軸上的差距，越偏離即代表樣本與文本按照齊夫定律應有的分布的比較上，顯得越特殊，我們說這即代表作爲樣本的關鍵詞越加關鍵。這樣的假定是否能夠成立呢？我們知道，《新青年》的思想變化反映了共產革命意識形態的確立過程，透過觀察偏離值的變化，可發現這正好反映了這種確立過程。

在圖四的盒裝圖中，會發現從第1卷到第9卷，每一卷的樣本與樣本之間的距離是不斷擴大的，也就是空隙越加明顯，這是因爲樣本的最大偏離值不斷地上升。第1卷到第5卷樣本的最大偏離值平均是13.84，到第6卷突然跳升到28.12，第7卷到第8卷維持在40以下，第9卷又再大幅度跳升到55.56，第10卷及第11卷則皆超過60。或許會有人質疑，即便不用看偏離值，只看詞頻高低，也能知道某些詞彙變得越來越加關鍵，然而不可忽略的是，這是因爲《新青年》各卷的文本大小差不多，若是文本大小差異很大，而不知其偏離值，即使某卷的一些詞的詞頻異常的高，其實有可能只是因爲其文本字數很多的關係，而不是因爲詞彙的關鍵性上升。簡單而言，觀察偏離值大小才是比較能夠確定詞彙關鍵的方法。

我們按照上述的觀察可根據樣本的最大偏離值來劃分三個時期：第1卷到第5卷樣本的最大偏離值平均是13.84，第6到第8卷是35.7，第9到第11卷是61.44。我們認爲，這正好反映了《新青年》的整個思想變化。

偏離值越大即表示關鍵詞越關鍵，第1卷到第5卷是社會主義尚未引入起得反響的時期，故此可知當時尚未有具體的、共識的觀念出現，因此出現的關鍵詞，自然偏離值偏低。

到了第6卷到第8卷之後，也就是1919年到1921年間，是社會主義引入取得反響的時期，而具有最大偏離值的關鍵詞，在第6卷到

airiti

第8卷分別是「社會」、「他們」、「他們」（見表一）。令人反思的是，除了「社會」之外，「他們」是常用詞，並不能說明意識形態逐漸成爲具體的、共識的觀念，然而亦不可忽略的是，第6卷到第8卷逐漸出現大量的屬於無產階級革命意識形態的關鍵詞，如「勞動」、「工人」、「資本」。這樣的落差可以得到這樣的解釋，此時期尚未出現觀念共識，因此《新青年》的作者們仍然沒有非常一致地共同使用屬於同一意識形態的概念語言，因此才用常用語言，即「他們」、「我們」來討論問題，故此最大偏離值的關鍵詞是「他們」。可以發現到，儘管《新青年》在第7卷中便明白表示從此接受社會主義：「現當第七卷開始，敢將全體社員的共同意見，明白宣佈」、「我們主張的是民眾運動社會改造」、「但對於一切擁護少數人私利或一階級利益，眼中沒有全社會幸福的政黨，永遠不忍加入」。[20]

但是概念的引進並不是立刻就能夠確立意識形態的，在此之後便出現了是否接受社會主義的討論，這可在〈關於社會主義的討論〉[21]一文中看出，文中收錄了許多作者與張東蓀（1886-1973）之間的爭論。張東蓀主張要先發展實業：「我此次旅行了幾個地方，雖未深入腹地，卻覺得救中國只有一條路，一言以蔽之：就是增加富力。而增加富力就是開發實業，因爲中國的唯一病症就是貧乏」，而支持社會主義的作者們卻認爲這是依賴資本家，他們想要的是階級革命，陳獨秀對他的反駁便是：

> 資本制度是制度不好，不是分子不好；政府和勞動階級不可靠，是分子不好，不是制度不好；分子不好可以改造，制度不好便要廢除了。諸君何以不想想法子努力改造政府或訓練勞動階級來施行新的生產制，而馬上便主張仍歸到資本家呢？[22]

---

[20] 〈本志宣言〉，《新青年》第7卷第1期，1919年12月1日。
[21] 陳獨秀：〈關於社會主義的討論〉，《新青年》第8卷第4期，1920年12月1日。
[22] 陳獨秀：〈關於社會主義的討論〉，《新青年》第8卷第4期，1920年12月1日。

airiti

到了第9卷到第11卷，我們知道中國共產黨成立後，標示著意識形態的確立，對照最大偏離值的關鍵詞正好能夠證明。第9卷到第11卷具有最大偏離值的關鍵詞分別是「主義」、「階級」、「主義」，都是意識形態用語，並且偏離值極高，表示「非常關鍵的關鍵詞」的出現，那正好印證了具體的、共識的觀念的確立。

可以這麼歸結，以偏離值來觀察關鍵詞的關鍵性是可以成立的，偏低的偏離值即表示越可能無具體、共識的觀念出現，偏高的偏離值即表示越可能具體、共識的觀念出現。而從偏離值的觀察會發現第1卷到第5卷，是「無共識觀念期」，第6卷到第8卷，是「思想討論期」，第9卷到第11卷，是「意識形態確立期」。這樣的分期與觀察關鍵詞詞數的變化，也是相互印證的。

（三）關鍵詞詞數與觀念確立的關係

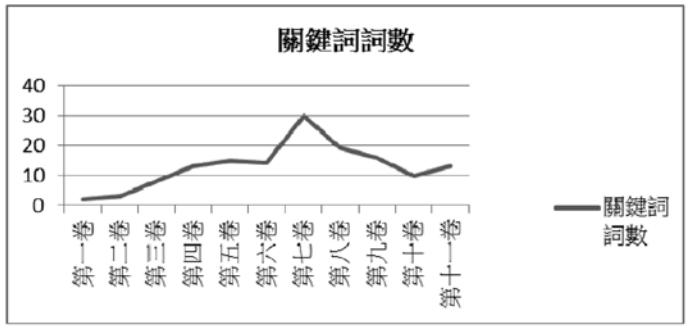
必須強調的是，引用齊夫定律偏離值計算於確立關鍵詞之後，關鍵詞的確立幾乎完全根據數據結果來確立，而非肉眼判斷，也非人力可控制。雖然因此能夠排除眾人主觀判斷的爭議，但也不得不生出這樣的疑問：「在完全沒有人文學者介入判斷的情況下，真的能夠囊括真正的關鍵詞嗎？」本文研究的出發點，一是要去除主觀判斷的爭議，一是要減少人文學者的負擔，我們從來就不否定人文主觀判斷的必要，在保留原始詞表（如附錄二）的情況下，人文學者大可自行斟酌。更加要小心的疑問是：「透過數位及統計方法所得的關鍵詞，是否具有代表性呢？」這一問題本文已經在第一小節的部分得到證實是具有代表性的。更深一層的是，「如果找出來的關鍵詞多，就代表該時期的重要觀念多嗎？如果少呢，又代表什麼？如果沒有，就意味沒有重要的觀念嗎？」

本文要強調的是，偏離值越大即表示越加關鍵，但不可反過來說，偏離值越小即表示越不關鍵。因爲我們的研究目的是要找尋「值

airiti

得注意的線索」，因此才如此重視偏離值的大小。在此之外，我們並不能因此而假定「除了那些明顯值得注意的線索，其餘皆不值得注意」。數位方法只能提供我們方便，是絕對不能因此排除文本中哪些部分重要或不重要。

本文要在此小節說明的是，關鍵詞詞數的多寡，是否能夠代表文本的某些意義。可以發現，從第1卷到第9卷，關鍵詞詞數呈現由少到多，再由多到少的現象，見圖六：



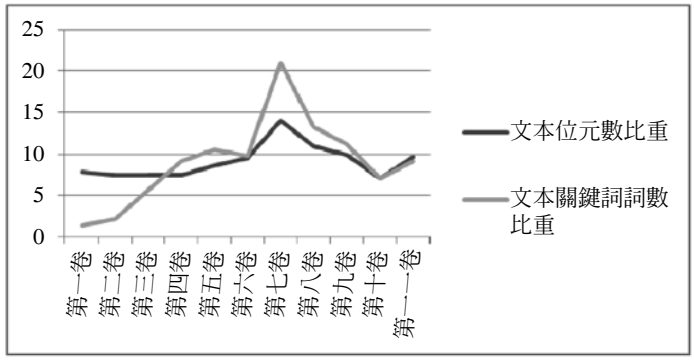
圖六：《新青年》十一卷每卷得出之關鍵詞詞數

我們從上節歸結出《新青年》的思想分期，對照關鍵詞詞數，可發現第1卷到第5卷的「無共識觀念期」的關鍵詞偏少，平均爲8.2個；第6卷到第8卷的「思想討論期」的關鍵詞偏多，平均爲21個；第9卷到第11卷的「意識形態確立期」的關鍵詞偏少，平均爲13個。我們知道，觀念濃縮於關鍵詞中，那麼關鍵詞的多寡是否能夠反映文本中，可以視爲明顯的觀念的多寡呢？從圖六的觀察可發現，這是能夠部分反映的。

可以這麼解釋，「無共識觀念期」的關鍵詞偏少，是因爲當時人們找不到核心價值來代表當時的觀念；「思想討論期」的關鍵詞偏多，是因爲思想激盪，人們各執一言，因此討論東西多，可以視爲明顯的觀念便多；「意識形態確立期」的關鍵詞偏少，是因爲人們有具

體、共識的觀念，觀念集中濃縮於某關鍵詞，故此關鍵詞就偏少。

然而，不能忽略的是，關鍵詞詞數的多寡，有可能是受文本字數多寡所影響，而與觀念變遷毫無關係。即文本字數越多，用 PAT-Tree 擷取的字串也就越多，自然有可能關鍵詞詞數也就會越多。將各卷的文本位元數比重與文本關鍵詞詞數比重[23]比較可發現，兩者升降的趨勢的確是有相關的（見圖七）。



圖七：《新青年》十一卷文本位元數比重與文本關鍵詞詞數比重的曲線圖

儘管如此，我們仍然有理由認爲關鍵詞詞數的多寡，能夠給予我們觀察觀念變遷的線索，因爲文本字數與關鍵詞詞數並非完全成正比關係的，這從圖七中便可以觀察到，文本位元數比重與文本關鍵詞詞比重並非是完全一致，實際上兩者的相關性只表現在第6卷之後。而且，我們以《新青年》十一卷作爲單一文本，其關鍵詞詞數也未必因爲字數大大增加而增多。全卷《新青年》的關鍵詞詞數一共有15個（見表三），與各卷相比，未見有太大的差異。

[23] 這裡所謂的比重，是指《新青年》的「單卷之量／十一卷總和」之間的比重。

airiti

表三：全卷《新青年》的關鍵詞

排名	關鍵詞
1.	主義
2.	社會
3.	我們
4.	階級
5.	他們
6.	不能
7.	一個
8.	沒有
9.	革命
10.	所以
11.	可以
12.	中國
13.	產階
14.	資本
15.	現在

再一次強調的是，我們利用數位元方法來處理文本，所得到的資料是作爲「考察歷史的線索」，而非「歷史證據」。在此前提下，我們說關鍵詞詞數的多寡與觀念變遷的關係，是指它提供了一個考察的線索，而非證據本身。無可否認從圖七可看出關鍵詞的多寡與文本字數多寡是相關的，然而若對照全卷《新青年》只有15個關鍵詞，又未必見得關鍵詞的多寡完全是受文本字數所影響，並且在閱讀文本之後，可以發現關鍵詞詞數的變化與觀念變化是相關的。故此，最保險的作法是，我們認爲不管是文本字數與觀念變遷，都是關鍵詞多寡的影響因素。在以資料作「線索」而非「證據」的前提下，我們認爲以關鍵詞多寡來考察觀念變遷，仍然是可行的。

airiti

五、結論

總結而言，利用齊夫定律偏離值計算應用於人文研究上，以《新青年》爲例，除證明了可行性之外，還取得三大成果：第一，能夠更加準確地將關鍵詞篩選出來，避免了主觀判斷的爭議，以及人工篩選的負擔，從而更加細緻地觀察思想變化，以及反映在文本上的語言現象，前者透過「國家」與「青年」這兩個關鍵詞來說明，後者透過「他們」、「我們」、「社會」、「主義」、「階級」、「產階」來說明；第二，印證了偏離值大小能夠代表詞的關鍵性，從而揭示出《新青年》的思想變化，從無共識觀念，到思想討論，最後到意識形態確立的過程；第三，印證了關鍵詞詞數能夠部分反映文本的思想變化，由於觀念濃縮於關鍵詞，關鍵詞的多寡能夠反映明顯觀念的多寡。

在揀選關鍵字時，此研究爲了研究上的便利，所以在每一卷都使用同一門檻值。不過，研究者可根據資料的特性分布，給予每卷不同的門檻值。這麼做或許能篩選掉一些不必要的詞彙，使得研究更有效率。不過，這裡必須注意的是，若卷數很多時，這樣做非常費神，弄巧成拙使得研究非常沒有效率。因此，如何取得其平衡點，是研究者必須去斟酌的地方。

判定離群值的部分，本次研究採用了 Tukey 的方式。其方法的優點是不受任何分配假設的影響，所以它適合使用在各種資料上。不過，如果偏離值的分布呈現爲常態分布，則可以使用其他建立在常態分布上的方法，如 Modified z-score（Iglewicz&Hoaglin，1993），再以一些離群值檢定（outlier test）爲輔，如 Grubb’s test（1969），這樣做，理論上可以得到更精確的門檻值。[24]

---

[24] 從圖四來看，《新青年》的偏離值大約傾向厚尾分布（heavy-tail distribution），且帶有一些右偏。故一般的常態假設所建立的檢定或許會不太合適。然而，我們試著使用 “Grubb’s test” 去檢定這筆資料，也發現有證據顯示這十一卷皆存在至少一個離群值。

airiti

我們使用數據尋找其歷史線索，必定多少會尋找到一些新的思考方向。不過，引用胡適先生的名言：「大膽假設，小心求證」，我們若找出某個新的歷史意義時必須小心，除非證據確鑿，否則不能過分肯定新的結果。在研究某個新理論時，學理上的探討、實驗、觀測資料的驗證是科學研究的三個重要的過程。很可惜的，歷史是不能被實驗的，巧合的是，在其他領域如醫學，也面臨了同樣的問題，如需知道某種藥物是否會造成潛在副作用，研究人員會基於道德倫理的原則，無法執行此實驗。在這種情況下，「統合分析」（Meta-analysis）這種研究方法開始發展起來，至今它在醫學發展中扮演了重要的角色，也解決了許多爭議。統合分析的精神在於結合各種小型研究，將各種沒法控制的不確定性因數排除，得出一個與類似實驗的結果。在文本研究上，大量的文本資料被數位化與資料化是未來必然的趨勢，所以大量的數位文本研究也必然會大量的出現，統合分析或許是一個有用的方法，我們相信它不但可以提供歷史學者一個宏觀面的看法，也能扮演歷史學者與數位人文學者之間的橋樑。相關的議題，金觀濤在〈數位人文的理論基礎〉一文中有更深入的討論。[25]

然而，本次研究的實驗之所以能夠被印證，是因爲《新青年》的思想變化過程非常明顯，可以說我們是在既有的人文研究成果上，來驗證齊夫定律偏離值計算是否能夠適用於人文研究上。我們所觀察到的三個成果，是否能夠普遍推廣到其他文本，是否能夠作爲一經驗法則，未來仍須在對照更多的文本之後，方能得出結論。

在此，我們可以提出一些值得繼續探討深究的問題，首先是本文只篩選出「二字詞」來展開研究，而對於那些超過二字的長詞段則未能篩選出來，無庸置疑的是，這肯定忽略了文本中一些關鍵詞的作用，例如在第9卷中，出現了三個關鍵詞「社會」、「階級」、「主義」，可以猜見的是，這三個關鍵詞，肯定會有兩種組合：「社會階

[25] 參見金觀濤：〈數位人文研究的理論基礎〉，收於項潔編：《數位人文研究的新視野：基礎與想像》（臺北：臺灣大學出版中心，2011年），頁45-61。

airiti

級」、「社會主義」，其中孰爲多寡，是對於文本的解釋很重要的。按目前只篩選出「二字詞」的技術來說，恐怕只能讓研究者自行從二字關鍵詞表中，觀察出其中重要的長字詞段關鍵詞，而不能讓數位技術直接篩選出來。對於這一點，除非使用的數位方法有自動辨識認字功能，恐怕要找出不同欄位的關鍵詞，只能透過多次設定欄位來展開多次篩選的工作。而其中不同字數的詞段又是否符合齊夫定律，亦是未來需要進行實驗的工作。

此外，本文所實驗的數位方法，其中一個問題意識是要避免不同研究者如何判定關鍵詞詞頻的主觀傾向，因此以齊夫定律的偏離標準確立出一個「標準值」，然而這對於欲對文本作出解釋的研究者來說，應該說是個「參考值」，因爲不同研究者根據其研究議題及學識，對於何者屬於「關鍵」的判定是不一樣的。那麼，這樣的「參考值」的意義，其功用不應是針對文本解釋的結論，而應是文本解釋的方便性，也就是能夠提供研究者一個觀察關鍵詞的範圍大小來作出取捨，而不至於陷入大海撈針的檢索裡。

以此而延伸的問題是，數位方法對於人文學者來說，到底是個怎樣的工具？本文認爲，數位方法所得出的數據，第一義應該是作爲一種「線索」，而非「證據」，不過根據此線索並透過嚴謹的文本分析，而能得出令人信服的結論的話，那麼數據就能變成一種「佐證」了。更進一步說，數位方法能夠提供給人文學者的，恐怕不只是一種圖求方便的工具，或者還可能作爲一種入手文本的觀測點，讓人文學者在數據所顯示的特異結果前，產生對文本解釋的啟發靈感，並且基於客觀的數據結果，而能夠提供一開放的資料平臺供大家討論。

最後，本文作爲一實驗性的文章，對於觀念的變遷，雖然企圖要作出某種的歷史解釋，但這不是蓋棺而定的結論，而是嘗試對歷史線索的「另一種梳理」，這是必須一再對讀者強調以免誤會的。

airiti

附錄一：《新青年》十一卷中最大偏離段的關鍵詞詞表

編號	第一卷			第二卷			第三卷			第四卷			第五卷			第六卷		
時間	1915年9月15日至1916年2月15日			1916年9月1日至1917年2月1日			1917年3月1日至1917年8月1日			1918年1月15日至1918年6月15日			1918年7月15日至1918年12月15日			1919年1月15日至1919年11月1日		
	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res
1	國家	526	15.01	人之	411	9.52	不能	496	12.23	先生	624	16.98	中國	788	15.45	社會	1243	28.12
2	政府	378	6.54	青年	398	9.04	社會	489	12.20	娜拉	609	16.55	我們	773	15.27	我們	974	17.38
3				不能	389	8.79	人之	477	11.85	不能	575	15.02	不能	737	14.08	所以	894	14.48
4							文學	450	10.58	文學	477	9.79	一個	701	12.84	一個	871	13.98
5							以爲	444	10.56	中國	458	9.03	所以	608	8.80	不能	856	13.79
6							不可	428	9.91	什麼	440	8.30	沒有	595	8.59	沒有	840	13.55
7							中國	391	7.94	沒有	419	7.38	可以	582	8.35	的人	818	13.02
8							吾人	339	5.00	我們	406	6.91	斯鐸	575	8.40	他們	740	10.00
9										一個	405	7.13	鐸曼	574	8.74	可以	693	8.31
10										以爲	400	7.11	他們	549	7.86	他的	670	7.65
11										可以	394	7.03	現在	534	7.46	主義	626	6.03
12										社會	369	5.78	先生	531	7.66	現在	625	6.33
13										夫人	366	5.85	文字	518	7.33	什麼	623	6.58
14													什麼	515	7.51	就是	622	6.86
15													的人	475	5.75			
16																		

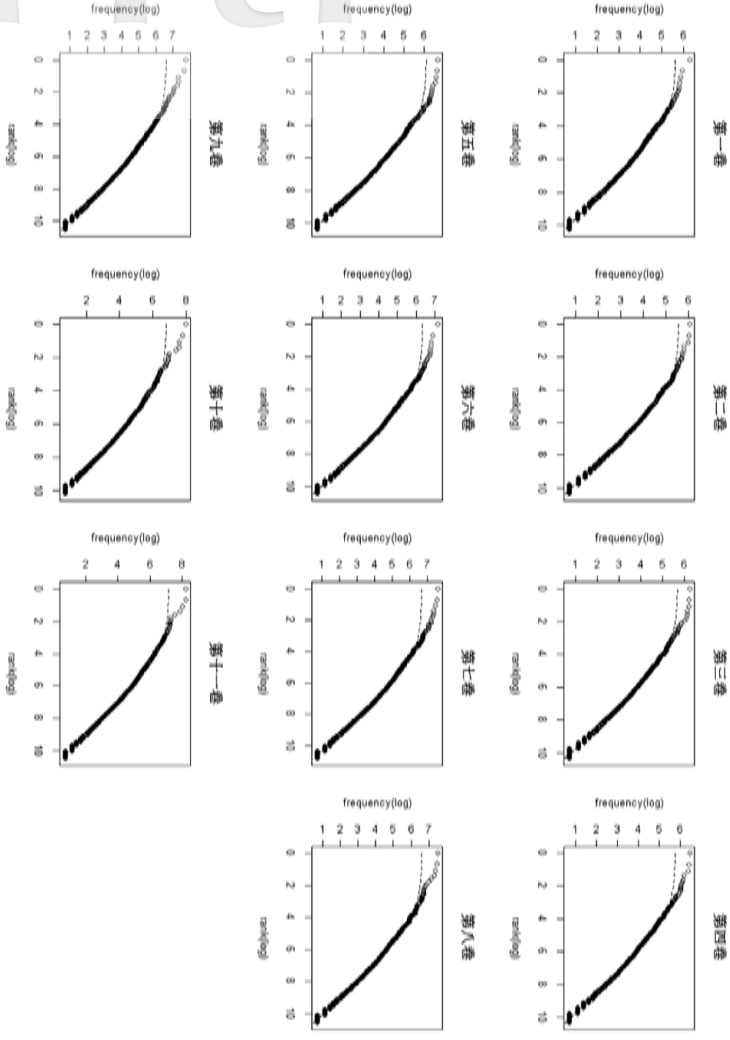
編號	第七卷						第八卷			第九卷			第十卷			第十一卷		
時間	1919年12月1日 至 1920年5月1日						1920年9月1日 至 1921年4月1日			1921年5月1日 至 1922年7月1日			1923年6月15日 至 1924年12月20日			1925年4月22日 至 1926年7月25日		
	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res	Word	Freq	Res
1	他們	1946	40.81	不是	938	9.71	他們	1759	38.18	主義	2231	55.56	階級	2866	64.60	主義	3607	64.17
2	我們	1872	38.69	生活	894	8.30	社會	1674	35.68	社會	1998	47.74	主義	2391	50.15	階級	3582	64.88
3	沒有	1720	33.75	問題	859	7.23	我們	1538	31.21	我們	1474	28.82	社會	1935	35.96	革命	2918	47.43
4	的人	1536	27.59	人口	853	7.27	主義	1496	30.22	階級	1472	29.39	產階	1912	36.24	產階	2515	37.07
5	社會	1521	27.48	們的	849	7.38	一個	1221	20.31	勞動	1364	25.89	革命	1548	24.66	工人	1862	19.22
6	一個	1499	27.11	時候	848	7.60	沒有	1081	15.44	他們	1127	17.35	資本	1078	9.09	資本	1450	8.03
7	勞動	1358	22.37	工作	835	7.35	可以	911	9.33	資本	1091	16.47	無產	1024	7.88	運動	1421	7.90
8	所以	1331	21.77	的時	826	7.25	現在	891	8.96	一個	1081	16.57	我們	1020	8.39	社會	1365	6.94
9	工人	1330	22.11	人的	824	7.42	勞動	847	7.63	沒有	997	13.76	共產	1006	8.54	我們	1362	7.52
10	可以	1179	16.88	這種	820	7.52	資本	815	6.76	不能	869	9.14	工人	883	5.26	無產	1349	7.79
11	就是	1165	16.71	的工	799	6.92	不能	812	7.01	所以	830	8.00				帝國	1284	6.45
12	不能	1158	16.79	的事	796	7.04	的人	803	7.01	就是	814	7.76				國主	1271	6.67
13	因爲	1145	16.63	但是	778	6.56	所以	799	7.21	可以	799	7.54				資產	1217	5.60
14	現在	1121	16.06	對於	767	6.35	政府	754	5.75	革命	740	5.52						
15	什麼	981	11.05	中國	739	5.44	因爲	741	5.55	會主	740	5.88						
16							對於	734	5.60	產階	699	5.21						
17							他的	719	5.31									
18							階級	710	5.25									
19							就是	698	5.06									

airiti

附錄二：《新青年》第1卷詞頻排名前一百的準詞表

序號	關鍵詞	詞頻	序號	關鍵詞	詞頻	序號	關鍵詞	詞頻	序號	關鍵詞	詞頻
1	國家	526	26	之所	180	51	國人	122	76	蘭西	97
2	政府	378	27	所謂	174	52	自然	122	77	內閣	97
3	人之	345	28	之人	168	53	希臘	116	78	思想	96
4	自由	333	29	精神	166	54	之自	115	79	一切	95
5	青年	316	30	所以	165	55	第一	114	80	子之	94
6	社會	315	31	家之	159	56	於是	114	81	宗教	94
7	不能	305	32	而不	159	57	爲人	113	82	其所	93
8	薩稜	290	33	意志	159	58	十二	112	83	將軍	93
9	夫人	287	34	人類	158	59	年之	109	84	之權	92
10	吾人	274	35	共和	154	60	十五	109	85	生之	92
11	不可	251	36	對於	150	61	爲之	108	86	協約	92
12	政治	250	37	不得	150	62	於此	108	87	文明	92
13	者也	246	38	英國	148	63	組織	108	88	無所	91
14	人民	231	39	今日	147	64	總統	106	89	之時	91
15	之事	229	40	吾國	147	65	墨子	104	90	之意	89
16	世界	225	41	教育	146	66	學者	104	91	國體	88
17	主義	224	42	仙瑪	145	67	之大	103	92	三十	87
18	國之	217	43	民之	142	68	不知	103	93	德國	87
19	革命	210	44	中國	138	69	可以	102	94	者之	83
20	姑娘	193	45	而已	136	70	獨立	98	95	之道	82
21	問題	190	46	人生	134	71	民族	98	96	言之	82
22	少年	188	47	日本	131	72	法蘭	98	97	會議	82
23	二十	185	48	之說	127	73	之中	98	98	關係	81
24	國民	184	49	權利	125	74	道德	98	99	行之	81
25	以爲	182	50	之一	123	75	人所	97	100	一日	81

附錄三：《新青年》十一卷的實際值與理論值



airiti

附錄四：本文統計方法的相關統計原理

首先，設 R 爲某字的排序，根據 Zipf-Mandelbrot 定律：

$$p(R=r)\ \alpha\ \frac{1}{(r+b)^a}, r=1,...,k\ \ ;\ a,b>0, \qquad (1)$$

因此，字數的機率爲：

$$p(R=r)=\frac{1}{c(r+b)^a},\ c=\sum_{r=1}^{k}\frac{1}{(r+b)^a} \qquad (2)$$

R 爲一服從 Zipf-Mandelbrot 定律之隨機變數。而 *p(R=r)* 爲一個離散的機率分配（Probability mass function）。其中，*r* 是字的排序，k 爲最大的字排序，*a,b* 爲該機率分配的參數。如果我們觀察式子 (2)，我們會發現，排序越後面的字，出現的機率會迅速下降。

接下來，由於 *a,b* 是未知的參數，我們必須找出一種方法去估計其數值。我們可將 *p(R=r)* 改寫爲：

$$\log(p(r))=c-a\log(r+b) \qquad (3)$$

從 (3) 可看出，[26] 字頻與字排序取對數後，爲一非線性的式子。我們進一步使用最小平方法（Least square estimator）去估計 *a,b* 。此作法與 Piqueira 等人 (1999) 所提出的方法非常相似。

$$\min_{a,b}\{\sum_{r=1}^{k}[\log(y_r/\sum_{r=1}^{k}y_r)\ -(\log c+a\log(r-b))]^2\}$$
$$\text{subject to}:\{a\ ,\ b\}>0 \qquad (4)$$

其中，y_r 爲該字排序爲 r 所觀察到的頻數。*a,b* 非線性最佳非線性解（Non-linear optimization）。其想法單純找出一組最佳數值去描述該筆資料。一般進階的數學或統計軟體如 R，Matlab 等，都會附有數值方法的求解程式。

當我們求出 *a,b* 後，我們便可以進一步建立 Zipf-Mandelbrot 的字

[26] 讀者可將 *p(R=r)* 看成 *p(r)*，此處使用兩種寫法只是爲了維持數學上的完整性。

頻理論值。設 $R^*=(R_1,...R_i,...R_k)$ 爲一隨機向量，R_i 爲每 i 排序的字所出現的字頻，假設 k,n 已知，R^* 爲一多項分佈（Multinomial distribution）。因此，對於任何 R_i，R_i 服從二項分配（Binomial distribution）分佈，也就是：

$$R_i \sim Bin(n, p(i)) \tag{6}$$

其分佈的期望值 (理論值) 爲：

$$E_{R_i} = np(i) = \frac{n}{c(i+b)^a} \tag{7}$$

而變異數爲：

$$Var(R_i) = np[1-p(i)]) \tag{8}$$

我們可進一步計算出標準化殘差（Standardized residual）：

$$e_i = \frac{R_i - \hat{E}_{R_i}}{\hat{Var}(R_i)} \tag{9}$$

所以，我們可計算標準化殘差 e_i，並把它定義成偏離值。若此數值越大，表示頻數越偏離 Zipf-Mandelbrot 定律，也就是我們所要尋找的關鍵詞。在實際計算上，n 爲總字數，而 k 爲最大的字排序。接下來我們將第 i 排序字的觀察字頻代入 R_i，$\hat{E}_{R_i}$ 與 $\hat{Var}(R_i)$ 的 a,b 則是代入式子 (4) 所算之值。其中 $\hat{E}_{R_i}$、$\hat{Var}(R_i)$ 爲理論值的相關點估計量。當全部的 e_i 算出來後，我們可畫出 e_i 的盒裝圖（Boxplot），進而判斷那些字爲關鍵字。若 e_i 越大，表示該字的頻率比預期次數多，很可能就是關鍵字。在字數很多時，e_i 的機率分配具有分配收斂 (Converge in distribution) 至標準常態的性質：

$$e_i = \frac{R_i - \hat{E}_{R_i}}{\hat{Var}(R_i)} \xrightarrow{d} N(0,1) \tag{10}$$

我們將此結果的理論證明放在附錄五。根據此結果，我們可將門檻設爲3,4 或更高。然而，由於這是大樣本之下的結果，在樣本不足，或 $p(i)$ 值很小時，e_i 不見得會很快收斂到標準常態分配。我們建議在使用上應配合盒裝圖加以判斷離群值。

airiti

附錄五：有關偏離值的理論證明

若 $Z_i \sim Bernoulli(p(i))$，則 $\sum Z_i = R_i \sim Bin(n, p(i))$。接下來，根據中央極限定理 (Central Limit Theorem)，下式會收斂到標準常態分配。

$$\frac{R_i - E_{R_i}}{np_i(1-p_i)} = \frac{\sum Z_i - np_i}{np_i(1-p_i)} = \frac{\bar{Z} - p_i}{p_i(1-p_i)} \xrightarrow{d} N(0,1)$$

又因為 $n\hat{p}_i$ 與 $n\hat{p}_i(1-\hat{p}_i)$ 具有一致性的性質 (Consistency)，因此：

$$e_i = \frac{R_i - n\hat{p}_i}{n\hat{p}_i(1-\hat{p}_i)} \xrightarrow{d} N(0,1)$$

airiti

徵引書目

項潔編：《數位人文研究的新視野：基礎與想像》，臺北：臺灣大學出版中心，2011 年。

臺灣大學數位人文中心編：《第三屆數位典藏與數位人文國際研討會論文集》，臺北：臺灣大學數位典藏研究發展中心，2011 年。

B. Iglewicz, Hoaglin. *How to detect and handle outliers*. ASQC Quality Press, 1993.

B. Mandelbrot, "Information Theory and Psycholinguistics." In *Scientific psychology*. Edited by B. B. Wolman and E. Nagel. New York: Basic Books, 1965.

Bruce M Hill, "Zipf's law and prior distributions for the composition of a population", *Journal of the American Statistical Association*, 65 (1970): 1220-1232.

Colin Martindale. SM Gusein-Zade, Dean McKenzie, and Mark Yu. Borodovsky. "Comparison of equations describing the ranked frequency distributions of graphemes and phonemes", *Journal of Quantitative Linguistics*, 3(2) (1996): 106-112.

DR Ridley. EA Gonzales. "Zipf's law extended to small samples of adult speech". *Percept. Mot. Skills*. 79 (1994): 153-154.

Frank Grubbs. "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, 11, no.1(1969): 1-21.

George A Miller. *The Science of Words*. Scientific American Library, a division of HPHLP. distributed by W.H. Freeman and Company. 1991.

Hang Xiao. "On the Applicability of Zipf's Law in Chinese Word Frequency Distribution," *Journal of Chinese Language and*

Computing. 18(1) (2008): 33-46

J. R. C. Piqueira, L. H. A. Monteiro, T. M. C. deMagalhães, R. T. Ramos, R. B. Sassi, E. G. Cruz. "Zipf's Law organizes a psychiatric ward," *J. Theoret. Biol*. no.198 (1999): 439-443.

John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley. 1977.

K. Z. George. *The Psychobiology of Language*. Boston. MA: Houghton Mifflin. 1935.

McCreight, Edward M. "A Space-Economical Suffix Tree Construction Algorithm". *Journal of the ACM* 23, no2 (1976): 262–272

R Rousseau, Qiaoqiao Zhang. "Zipf's data on the frequency of Chinese words revisited". *Scientometrics*. 24(2) (1992): 201-220.

S Naranan. VK Balasubrahmanyan. "Models for power law relations in linguistics and information science". *Journal of Quantitative Linguistics*. 5 (1998): 35-61.