數字人文及其史前史

〔美〕戴安德(撰)　林太平(譯)

人文學科學術研究越來越多地得到各種形式的計算分析的協助,包括自然語言處理和文本挖掘、網絡映射及地理測繪等等。將人文學科與計算技術聯合在一起的是它們在數據收集、處理及分析當中的基礎。研究數據即是找出模式和反常事例:突出的內容也許揭示出關於歷史關係或審美趨勢的新事物,或確定了已被接受的知識。好的分析始於好的數據:語料庫或數據庫的分佈決定了研究者能提出的問題的類型和收到的答案的質量。對於任何有基礎統計學訓練的人來說,這一說法都不意外,取樣是統計學這一學科不可或缺的部分。這也包括社會科學,它長期以來充滿了定量分析帶來的自負和狂妄(馬克思、弗洛伊德和韋伯這樣偏重理論思考的社會科學學者,現在似乎被放在歷史和文學課程中閱讀了)。但是在人文學科,尤其是文學和歷史領域,數據被賦予的新的重要性,這已經成為關係到本領域發展方向的持續且極為重要的對話的一部分。

蹤跡:人文學科與定量

數據驅動的定量分析,是否從根本上不同於更古老的專注細致的閱讀實踐和運用證據對歷史的重構?這一問題是開放性的,但回答起來風險頗高。沒有正確或最終的答案——任何做出回答的人都在冒風險,可能忽略數據為人文學科知識生產帶來的真正收獲或真正威脅。一個人對這個問題的回應,或許可以代表他對正在國內外大學院系和研究中心普及的"數字人文"領域的態度。

隨著數據以及從數據化到定量尺度分析的各種過程相應變得更為盛行,我們面臨著追溯數據在本領域較早期的表現和使用的任務(當然,更普遍地考察數據在現代存在和知識的發展中的作用也很重要)。

數據分析和模式識別的早期形式是何時、何處、如何變成不可或缺的(或僅僅是可能的)人文學科研究工具的?對這一問題的回答揭示了被遺忘或被壓制的歷史插曲,前幾代學者們曾經運用數據分析來閱讀、思考並最終產出新的文化和歷史知識形式。因此,一部數字人文的史前史會提供有助於我們探討數字人文現狀的重要語境。這不僅包括這一新領域無數引人注目之處,還包括其局限、挫折、失敗,以及它們引起的抵抗和懷疑主義的形式。也許,這樣的研究甚至能修正對本領域更廣闊的自我認同,挑戰長期以來關於人文學科和定量學科間分裂的假設。

"史前史"這個概念本身是有問題的。這個詞讓人注意到兩個時代間的一道溝壑,似乎因為某些決定性事物或特點的缺乏,而需要由較後的時期時代顛倒地界定較早的時期。將兩個時代綁在一起所暗示的不只是主題上的關聯,還有關於連續性、甚至因果性的斷言:因此史前史在擴展一個事件視界的同時,既肯定、又挑戰了傳統歷史分期,並指向一個更早的起源。

談到數字人文,這個領域儘管還相對處於初創期,卻已經非常發散且眾所周知地難以精確定

義,要撰寫一部本領域的連貫歷史已相當困難——更別提發掘史前史了!如果僅僅關注相對固定的標志,如"數字人文"這個名稱(或影響較小但更惹人激動的名稱,如"遠讀"或"文化分析學"),單純追蹤它在學術話語中的出現,那不過是一種有限的話語分析形式。它並未告訴我們該領域中的學者借此實現其自我意識的那個過程。而過於關注詞語的致命錯誤在於,歷史的施動者往往缺乏術語去描述其存在狀況(例如,只用想想氧氣這個詞,人類在識別它之前已經呼吸氧氣很久了)。

要突破一個名稱的認知論邊界,我們就會遇到本體論問題。"數字人文"這個歷史對象到底是什麼?其"數字"是否只應用於數字計算機?如果這樣,那麼對其歷史範圍的追溯不會早過 1950 年代,如意大利耶穌會會士羅伯托·布薩(Roberto Busa)和計算機公司 IBM 合作,做出托馬斯·阿奎那作品的詞語索引,全都記錄在打孔卡片上;或者 1964 年 IBM 公司組織的會議,討論"人文學科計算"和文學數據處理,這一會議開啟了一段富有創新的實驗和對話的時期,包括 1966 年創建《計算機和人文學科》(*Computers and the Humanities*)雜誌。要更充分地理解 1970 年代晚期和 1980 年代整合之前的數字人文,必須仰賴更多的研究。

這樣的研究已表明,在人文學科變得"數字"之前幾十年,學者們已然在用計算機做實驗,探索與文本、作者身份及語言相關的問題。但這些例子遠未窮盡可能的數字人文史前史。如果不是通過計算機的作用,而是通過方法和技術,採取更為全面的路徑去界定數字人文,就能確認許多跨越計算年代和前一個世紀之間的關鍵性關聯。實際上從 19 世紀早期以來,計數和計算能力的各種形式就已是人文知識生產的一部分——那段時期中,人文學科本身變為我們如今認可的模樣。例證涵蓋 19 世紀德國語文學家為古典希臘詩歌格律計數,以及物理學家托馬斯·門登霍爾(Thomas Mendenhall)為確定某些莎士比亞作品作者身份所做的根據字母數量測量詞語的實驗。和早期計算時代的歷史一樣,這幅畫面仍在填充之中。這些早期歷史最有意思的事例裡,有一個是 1920 年代定量分析和統計推理在清華大學人文學科的形成和制度化中所起的作用。

模式:觀看的距離

這則插曲的發端,是 1922 年 11 月梁啟超在東南大學做的一場講演。講演被抄錄下來並發表在讀者頗眾的《晨報副刊》上,給我們一個令人著迷的——可惜也是被忽視的——關於一項發明的記錄。在講演中,梁啟超介紹了他正在開發的一種新方法,命名為"統計歷史學"。

如其名稱所示,這種方法將統計學原則應用於歷史數據,以便確認歷史潮流和模式。梁啟超的靈感來自中國歷代人口的起落。對歷史人口的興趣並非新鮮事,數十年來已經引起對社會改革感興趣的晚清知識分子的密切關注。人口確實位於對馬爾薩斯和優生學等"生命權力"不斷增長的興趣的中心;1903 年梁啟超在他所辦的《新民叢報》上發表過一篇關於人口的文章。在這篇較早的作品中,他考察近代歷史,解釋並批判清政府統計數字的不可靠和國家對人口的管理不善。但 20 年後,梁啟超顛倒了統計學和史學的關係。此時的梁啟超沒有用歷史去解釋一個流行的統計數字(中國人口 4 億)及其對中國國內及國際形勢的涵義,而是尋求使統計學為寫作歷史服務。(這不代表從政治或當代的重要性撤退,而代表梁啟超對學術嚴謹的興趣,是一種轉向,體現於他在其事業晚期產出的雄心勃勃的學術著作之中)。簡單地說,新方法意在收集並評估即使是非常細致的學者也可能在閱讀歷史記敘時忽略的所有那些小的細節和事實。梁啟超表述如下,令人難忘:

> 欲知歷史真相,決不能單看台面上幾個大人物幾樁大事件便算完結;最要的是看出全個社會的活動變化。全個社會的活動變化,要集積起來比較一番才能看見。往往有很小

> 的事,平常人絕不注意者,一旦把他同類全搜集起來,分別部居一研究,便可以發現出極新奇的現象而且發明出極有價值的原則……統計學的作用,是要"觀其大較"。換句話說:是專要看各種事物的平均狀況,拉勻了算總帳。

一個世紀後回頭看,梁啟超持續以其興趣之廣博和智識之創新令我們驚詫。"統計歷史學"是典型的現代時刻,反映出民國初期學者對以新的科學方法研究歷史和文學的興趣激增,其社會學傾向反映了當時特有的"對事實的激情"。(參見 Tong Lam, *A Passion for Facts: Social Surveys and the Construction of the Chinese Nation-State, 1900-1949*, Berkeley: University of California Press, 2011)但梁啟超的方法作為後來年鑒學派和計量史學定量分析實驗的先驅尤其突出。甚至可以將其看作數字人文的先行者,特別是對於弗朗哥·莫雷蒂的"遠讀"及其對"比文本小得多或大得多的單元:裝置、主題、轉義——或體裁和體系"的關注而言。(參見 Conjectures on World Literature, in *New Left Review*, 1: January-February, 2000, Online at: http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature, accessed 4/15/2017) 其實,三年前我和姜文濤在《山東社會科學》開創"數字人文"專欄時,便決定要採納梁啟超傳神的"觀其大較"的說法作為名稱,向這一遙相對應的情形致敬。

但是,"統計歷史學"開啟了數字人文的史前史的同時,也讓人對中國更古老的史學傳統產生新的認識。具體而言,梁啟超不僅將其發明歸功於西方統計科學被引入中國。他將其定位為對現代西方科學和清代考證學學者的綜合,如顧棟高及其權威研究《春秋大事表》,將《春秋》"折碎"為一系列表,將姓氏、事件和地點組織為條理分明的表冊。顧棟高著作對梁啟超的影響顯示了確定"史前史"自身位置時的偶然性和棘手。即使將《春秋大事表》看作一種起點,我們也必須承認其主要技術、也就是"表"本身就有一段可以追溯至清代之前很早的史前史。"表"作為給信息分組並形成結構的方式,讓信息可獲取,讓一個集當中的各點易於比較,實際類似於一種早期的數據庫形式,或者數據框架。梁啟超方法的新穎之處在於用數字象徵性地處理數據(儘管用數字指代生活和事件這個做法本身就很古老,可追溯至有文字記載的歷史之前很久。在英語中,digit 一詞既是"數字"也是"手指",顯示了後者作為編號索引的用途。人類——和文化——一直是數字的)。因此,歷史統計學是一種原初的事件,但又建立在更早的實踐和技術之上。

在 1920 年代中國的語境下,梁啟超更感興趣的是將來的學術研究,他對此滿懷期待。比如,他預想了一個大規模的二十四史通表項目,準備以此補充中國的二十四史。儘管他在啟動這個雄心勃勃的項目之前就去世了,但他的方法在同時代人中產生了廣泛影響,在其後十年裡啟發了許多對歷史人物地理分佈的研究(此處我們可以提出另一個當代學術研究的對應事例,即哈佛大學的中國歷代人物傳記資料庫項目,展示了非常豐富的中國文獻中的群體傳記信息)。梁啟超的影響在古典學者衛聚賢(1899~1989)的著作中最為顯著,後者直接將歷史統計學發展成為大家都可以運用的通用方法。由於衛聚賢有興趣超越歷史社會學、進入文本分析領域,對於數字人文的史前史而言,他的事例尤其有趣。

工具:用算盤做歷史研究

1920 年代中期,梁啟超和陳寅恪、王國維等其他聲名卓著的史學家在清華大學國學院形成了一批核心導師,這個機構儘管短命、卻生機勃勃,在中國現代學術知識發展中起到關鍵作用。這一時期和我們現在很相似,學術機構和學術性學科變動很大,但也提醒我們,最有意思的一些觀點來

自於調和傳統和現代認知論及方法的嚴肅嘗試。

回頭看,儘管國學這個學科確實逐漸被看作保守的學術研究領域,部分原因在於其本土主義,以及反對(或者至少不同於)五四的知性主義的世界主義。但依然值得回想起,與國學相關的人當中,有許多明確力圖以新的、現代的工具探索中國歷史。這一項目最著名的倡導者是胡適,他同傅斯年和顧頡剛一道,呼籲"重新整理"傳統史學("整理國故"),使其更接近自然科學的原則。眾所周知,胡適敦促學者們"大膽的假設,小心的求證"。然而,在這次學術方法科學化的行動中,惟有梁啟超的歷史方法應用了統計科學,將論證建立在對一般狀況的計算基礎上,而非建立於對邏輯不一致的辨別上。但在國學院內部,歷史統計學相對邊緣化,似乎並未教授給學生,梁啟超的同事們也未在研究中採納這個方法。

國學院裡一名叫做衛聚賢的學生的研究是個例外。衛聚賢進入國學院時,教育背景有點不那麼傳統,轉入歷史之前,他曾在商業學校學過會計。後來他講述過他如何頻繁地被清華同窗揶揄,後者看到他手持一把算盤做研究,便嘲笑他是沒受過教育的"商人"。但衛聚賢對會計和數據報表的興趣讓他尤其被統計學的實證主義吸引。他在清華期間努力工作,將梁啟超的方法擴展為一套更完滿的工具,發表了一系列文章解釋如何用統計學研究過往,並展示了這些工作的結果。

要清楚了解衛聚賢的計劃,只需瀏覽 1929 年出現的一篇關鍵文章,這是衛聚賢名為《應用統計的方法整理國學》的主要研究內容的縮略版。該文發表於《東方雜誌》,這是當時傳播最廣的流行雜誌之一,可見衛聚賢和雜誌編輯展望著歷史統計學會有廣泛吸引力;這篇文章也有助於加強衛聚賢作為此方法首要鼓吹者的名聲。在衛聚賢手裡,"歷史統計學"的應用延伸至"統計歷史學",其中,任何文本都能變成某種獨立的詞或字的個數,所有這些反過來又能被計數、分析。文章將統計方法的價值牢牢固定在圖表等數據視覺化的修辭和視覺吸引力中,有超過一打制作精細的漂亮餅圖、圖表和其他可視化圖,用於比較《春秋》和《左傳》的語言及內容。

幾年後,衛聚賢在上海持志學院做了一系列講演之後將文章擴展,於 1934 年出版了一本課本,就叫《歷史統計學》,目標是更充分地讓這種方法可操作。衛聚賢的著作清楚說明了"數據"的定義、如何從文本中獲取數據、如何計算並視覺地表現,以便對史實做出推斷,這本書是對歷史統計學最全面的解釋和演示。在此處,歷史學家被重新想象為一位要勘測過往的社會科學家。衛聚賢綜述了幾種能得到數據的觀察方法,例如直接測量或取樣,他提出一種新的範疇,即"索隱"(indexing,也作"引得"),指的是直接從歷史文獻中提取數據的過程。衛聚賢依靠自己在清華的經驗,詳細描述了索隱的物質性及腦力勞動,諸如指導讀者避免一邊讀文本一邊做標記,因為他警告說大腦無法同時做這兩件事;或者指導讀者用鋼筆或彩色鉛筆標記地名、事件、人物或章節,之後收集到索隱卡上(他為索隱卡的格式提供了很方便的模板)。

這一討論本身便是令人驚歎且原創的對數據化的敘述。而這個敘述又被納入對統計方法更廣的描述中,跨越三個分析階段,讓文本數據越來越抽象或經過處理:開始的時候用統計譜來處理文本,接著用數字表達數據,將其轉為統計表,能從統計上分析,最後用一張統計圖總結分析結果,讓它們容易理解、視覺上吸引人。有了這一套方法,任何人都可以從事統計歷史學。但尤為令人驚訝的是這一過程如何恰切地描述了現在的數字人文。中國任何關於這個主題的入門課程都能用這套異乎尋常的教科書作為第一周的教材。但是要解釋、闡明這一史學方法則需要三分之一的學習時間。課本其餘部分也值得在我們對史前史的批判中提及,因為這些部分是關於"中國統計學史"的。衛聚賢在這裡顯示出他敏銳意識到需要以本土主義詞匯和民族史來為他的方法措辭。

另外,如導論所言:“中國人的保守觀念傳統思想非常的大,以為統計學乃是外來的,中國的國學用不著用外人的方法去研究。殊不知統計學是中國的土產,中國的古人曾屢為用;現在將中國土產的圖譜學略為改造為統計學,使之研究中國的國學,當較前人的成績為佳。故作此中國統計學史一文,以為呼醒!”這不僅僅是為了讓他的方法在同儕中獲得合法性而採用的諷刺性戰略:衛聚賢很誠懇地試圖證明許多統計實踐在中國出現要早於歐洲。相比有傾向性地宣稱統計學“起源於”中國,更有趣的是衛聚賢廣泛探討了中國歷史上信息管理的諸多種類。其結果就是前所未有的中國數據實踐研究的歷史,從我們今天的有利眼光看來,它構成了數字人文史前史的史前史,是一種無窮盡的分層,挑戰了認為當代數字分析是獨特的或無前例的觀點。

結論:一朵忽然之間綻開的花

衛聚賢和梁啟超一樣,對歷史統計學發展的追求並未超越最初的投入。值得注意的不是這些學者嘗試的結果,而是他們建議採用的那些方法。這兩位學者共同留給我們一段引人入勝的插曲,它不僅給現代中國學術研究中的信息管理和數據分析的更具系統性的歷史提出了可能性,還提供了關鍵的比較點,憑借它可以考察我們當下的時刻。歷史統計學和數字人文都想要將實證或定量方法同一個傳統上更具闡釋性的知識領域相結合,在此範圍內,我們已指出二者間一些較為顯著的相似之處。

但差異也同樣重要。儘管缺失的似乎只是計算機和現代人機界面促成的勞動自動化——制表、數據提取和數學分析,但這些技術也確實造成了分析規模及複雜性上的顯著差異。歷史統計學確實能夠在一把算盤上操作。相比之下,有大量計算及反複的過程,例如恰當建立的主題建模技術(讓人得以同時分析數百萬文檔,並根據共享主題或“話題”在數據集內識別分組),就反映出在數據中辨別模式的能力的重大飛躍。主題建模這種技術實際非常複雜,以至於形成某種黑箱,將分析處理與人類操作者隔離,讓計算機及其算法成為主動的夥伴而非僅僅是被動的設備。這段時期另一個關鍵差異是態度上的。梁啟超和衛聚賢都是實證主義者,將定量知識視作某種確定性而接受。換句話說,統計歷史學代表了對系統、理性化、效率和進步的特別現代的熱情。如果其倡導者發現它有缺陷,那只是因為當時的技術不足以匹配其遠見。(這是真的,儘管這種方法和較早的學術研究模式有關——考慮到考證學者對實證知識和文本真實性的興趣,很容易想象他們欣然接受梁啟超和衛聚賢的實證主義態度。)數字人文學術研究的最佳案例對於其結果的局限性持開放和反省的態度,包括項目設計和統計顯著性及置信度方面。

這些差異和相似能夠共同闡明今日數字人文的歷史獨特性。某種程度上,這一插曲表明,數字人文並非全然是衍生物或完全是新近的舶來品。當然,我的意思也不是說數字人文在某種意義上起源於中國,我已經強調過,所有起源故事都是成問題的。相反,這個插曲是關涉全局的更大拼圖中的一小片,這幅拼圖由松散的片段組成,很大程度上並不連貫。但我們無需繪制歷史統計學和數字人文之間的直接線性關係,亦能從中獲取靈感。日本學者柄谷行人描繪夏目漱石的作品如“一朵未到季節便已開放的花,因此沒有留下種子”(一朵忽然之間綻開的花),這說法令人難忘。對於1920年代的清華學者,我們也可以這麼說。不過這朵花的季節已經到來。數字人文如今正在成長,讓我們自許為這種早期探索和開放精神的繼承人。

(作者 Anatoly Detwyler,係威斯康辛大學麥迪遜分校亞洲語言與文化系助理教授)

[責任編輯　桑　海]