運用文字探勘技術探討性相關議題之研究
——以PTT論壇 feminine_sex板為例

余采蓓
樹德科技大學人類性學研究所
施俊名
樹德科技大學人類性學研究所
郭洪國雄*
樹德科技大學人類性學研究所

摘 要

隨著全球進入資料科學的時代，巨量資料來源不僅僅只有結構的資料，文字及不具結構化的資料在我們的生活中也到處可見。使用網路蒐集資訊儼然成為上網的重要目的，挖掘民眾所關注之性相關議題便成為瞭解民眾對性的態度及性知識是相當重要的方法。本研究使用R語言撰寫爬蟲程式來自動抓取批踢踢（PTT）論壇女性性板（feminine_ sex）的文章，蒐集一個年度共1,438篇的文章，從語料庫大量的文字資訊中，我們其實很有機會在性議題裡發展出各種有潛力及有趣的應用，這正是本研究在文字探勘技術的目標。feminine_ sex板經過自然語言斷詞處理，研究結果顯示出現次數最頻繁的前三個詞彙為醫生、問題與男友。主題模型透過K-Means集群演算法，分析結果經命名後呈現大眾討論的議題大多圍繞在親密關係、避孕諮詢以及衛生醫療三個主要議題，而此研究結果亦可提供教育及醫療相關單位，實施性教育及衛教訓練的補強。

關鍵字：文字探勘、性議題、批踢踢實業坊、網路爬蟲、主題模型

DOI:10.6206/SIS.201901_9(2).0004
投稿日期：2018.11.10；接受日期：2018.12.16
* 通訊作者：郭洪國雄；E-mail: kuoshung@stu.edu.tw

Ainosco Press

Applying Text Mining Techniques to Sexual Issues on PTT feminine_sex

Cai-Pei Yu
Graduate School of Human Sexuality, Shu-Te University
Chun-Ming Shih
Graduate School of Human Sexuality, Shu-Te University
Kuo-Hsiung Kuo Hung*
Graduate School of Human Sexuality, Shu-Te University

Abstract

Entering the era of information science globally, we find that big data not only contain structured information but also include text and unstructured information. The use of the internet for information collection has become one of the important purposes of the internet. Therefore, it is very important that doing research on how people concerned about the sexual issues could help us to understand people's attitude on sex and their sexual knowledge. This study used the web crawler which created by R language to automatically extract the articles from the feminine sex board, collecting a total of 1,438 articles in one year. Then, from a large amount of information in the text corpus, we were actually given a chance to develop a variety of potential and interesting applications in sexual issues, which is the purpose of this study in the text mining techniques. After the word segmentation in the natural languages processing, the results showed that the three most frequent words in feminine_sex board are doctor, problems, and boyfriend. We used the K-Means cluster algorithm on the topic model. After classifying the analysis results, we get to know that the public discussion topics are mostly about three main issues, which are the intimate relationship, contraceptive counseling, and health care. Hence, we can

DOI:10.6206/SIS.201901_9(2).0004
Manuscript received: 2018.11.10; Accepted: 2018.12.16
* Corresponding author: Kuo-Hsiung Kuo Hung; E-mail: kuoshung@stu.edu.tw

Ainosco Press

provide the results for the respective educational and medical authorities to advocate sex education and to improve health care training on this related topic.

Keywords: Text Mining, Sexual Issues, PTT, Web Crawler, Topic Model

壹、緒論

一、研究背景

隨著網絡科技的發展，民眾的資訊交流與網際網絡的關係日益密切。根據2017年台灣寬頻網路使用調查報告，高達83.4%的民眾有使用網路，其中25 ~ 29歲使用率高達98.1%，高於其他年齡層；研究所及以上學歷之使用率為98.2%，高於其他學歷（鄭天澤、陳麗霞、楊亨利、胡正文、鄭閔安，2017）。網際網路盛行的原因是它開闢了各種交流空間和形式，包括論壇、部落格、聊天工具等，民眾便可以進行即時或者非即時的交流，更方便且有效率地交換資訊和溝通，從鄭天澤、楊亨利、陳麗霞、胡正文與劉千鳳（2015）調查發現，網路使用者所從事的活動在「獲取資訊」項目上以瀏覽網頁（world, wide, web, WWW）的40.0%比例最高，查詢新聞氣象（21.3%）居次；「文字溝通」部分是以「社群網絡」（如Facebook、噗浪、Blog、Linkedin、Twitter、Instagram等）的60.1%比例最高，其次則是使用即時通訊軟體（如Line、Skype、WhatsApp、WeChat、Juiker、Yahoo、Messenger）有56.3%。第三高的則是使用「電子布告欄」（如BBS、PTT）有0.5%。許多使用者已習慣在網路交換或取得所需之資訊，訊息內容包羅萬象，從健康、旅遊、工作、財經、情感、交友、時尚、學術、娛樂、宗教、藝術、運動、醫療等各面向，其中性相關議題更是日增月盛。利用網路交流性議題的動機可能是網際網路突破了面對面交流時的限制，加上其匿名性與安全性，也可能讓使用者感到更有自信且強大，在網路上分享性幻想、性行為等會較少顧忌。Cooper、Delmonico與Burg（2000）也發現相較於色情場所，女性更偏愛於網路聊天，因為網路可以提供安全的平台讓人暢所欲言。

性和日常生活息息相關，但在過去傳統及威權下的台灣，關於性的文學、藝術、探討、言詞等，都受到貶抑（古鐘响，2009）。人們所使用與性有關的文字或討論之議題，反映了歷史脈絡上當時人類社會所關注或討論的現象（Plaud, Gaither, & Weller, 1998）。有關性意涵的文字使用也會出現變化，很多辭彙都漸漸地演變成普遍用詞，反映了社會價值的變遷（Sanders, 1978）。針對Google搜尋引擎進行的大數據分析研究指出「無性婚姻」的搜索次數是「不幸婚姻」的3.5倍，是「無愛婚姻」的8倍，顯示在婚姻中，人們的性生活並沒有那麼活躍，對於婚姻的最大不滿或抱怨是沒有性生活，已婚者抱怨配偶不願做愛的次數，遠超過抱怨對方不願交談次數的16倍；研究還指出即使是沒有結婚的伴侶，也會容易頻繁地抱怨自己缺乏性生活，搜索「無性交往關係」

的頻次，僅次於「不良交往關係」（Stephens-Davidowitz, 2015）。Moreira等人（2005）利用隨機抽樣的電話採訪和問卷調查了29個國家27,500位男女性，年齡介乎於40 ~ 80歲，探詢過去一年中是否曾就性相關問題進行求助或尋求解答。他們發現大多數的受訪者表示至少有一個以上的性相關疑問，然而高達77.8%男性與78%女性未有尋求專業的協助或建議，有17%男性及16%女性會透過網路及報章雜誌尋求協助及答案；調查結果也顯示隨著年紀的增長，主動去心理諮商或婚姻顧問及詢問家人朋友的頻率逐漸下降；相較於年長者，年輕人對於性相關問題會更主動尋求專業協助或是詢問親友之意見與經驗。由此可知，性是在婚姻或伴侶關係中重要的元素，大多數人對於性相關議題都有疑惑，但並未透過專業的管道尋求協助，而選擇自行瀏覽網路或書籍。

利用網際網路討論性議題是當今社會的普遍行為，性相關資訊、問題或討論相當多元且繁雜，而且訊息量如恆河沙數，可以作為另一個出發點來探究大眾對性議題的認知與看法。現今許多資訊都是以文字的形式保留於網路之中，文字探勘（text mining）技術可以藉此透過自然語言方式處理非結構或半結構化的資料，利用文字量化後的特性，找出高頻率關鍵詞之間的關聯性，達到大數據資料分析的目地，就是挖掘出巨量文字資料中所隱藏的規則與結構。

二、文字探勘技術

許多公司積極發展大數據，利用社群軟體的貼文，分析市場反應，產品服務及未來策略，即使是曇花一現的訊息，都可能創造出新市場或產品的重要資訊，故這些公司設有實驗室隨時掌握全球資訊脈動，隨時蒐集反饋，即刻處理與回應消費者，或開始發展新產品及市場（George, Haas, & Pentland, 2014）。Mishra（2016）指出資料探勘能整合不斷增加的數據或文字，像零售商店中的產品條碼、無線射頻辨識（radio frequency identification, RFID）標籤、不斷變化的天氣、Facebook發文與電視的收視率等，加以分析不僅能提供關於數據結構之有意義的信息，也能運用於未來的發展活動。大多數企業運用資料探勘在發現知識、資料視覺化及修正資料，並利用資訊科技系統，自動進一步的篩選，讓使用者搜尋到符合需求之資料，減少垃圾訊息的干擾（Nicholson, 2006）。在商業利益的推波助瀾之下，各種大數據研究和解讀電子訊息的需求陡升，興起了資料探勘（data mining）熱潮，也進一步延伸出文字探勘的應用。

所謂文字探勘是針對所蒐集的特定巨量文字，進行編輯、組織與分析的過程，以發現其間隱含的關聯特徵或有趣新穎的模式，為分析師或決策者提供有效且關鍵的訊息（Blake, 2011; Sullivan, 2001）；具體過程主要以非結構或是

半結構化的文件資料進行探勘，挖掘出文字資料詞語模型中所隱藏的規則與結構，從巨量資料中以自動或半自動化的方式，識別出共同且經常出現的術語和關鍵字，然後探索和分析其關聯性（Aggarwal, 2015; Berry & Linof, 1997）。文字探勘不只是一種技術，而是結合多項專業技術的研究，利用自然語言處理工具在非結構化文件中進行文字樣版（patterns）萃取工作，從文件中自動選取未知且有用的隱藏資訊，有利於知識探索的發展（Adriaans & Zantinge, 1996; Delen & Crossland, 2008）。

關於探勘方式，國內外專家學者各有不同的研究步驟與分類，Han與Kamber（2001）將資料探勘方法分為六大類，分別為關聯分析（association analysis）、分類（classification）、預測（prediction）、集群分析（clustering analysis）、搜索（search）和最佳化（optimization）；黃文與王正林（2015）則劃分為分類、分群、關聯性規則、異常和趨勢發現等四部分。Mishra（2016）整理關於數據挖掘的過程及步驟：（一）從數據和資料庫提取所需的數據；（二）對數據執行完整性進行檢查，刪除多餘的字和不相關信息；（三）結合各種其他不相交的信息，組合成數據庫；（四）數據轉換技術，並檢視需要幾個屬性與特徵；（五）輸入識別的特徵值；（六）知識呈現，形成可視化資料庫。

文字探勘常用之技術與演算法有許多種類，如類神經網路（artificial neural network）、決策樹（decision tree）、基因演算法（genetic algorithm）、規則推論法（rule learning）、單純貝氏演算法（nave Bayesian model）以及模糊理論（fuzzy logic）等（曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯，2005；黃文、王正林，2015；Adriaans & Zantinge, 1996）。要應用何種技術往往視乎對象，呈現的結果也會有很大的差異。

文字探勘技術能廣泛應用於各領域之中，國外研究針對網路訂房之飯店顧客滿意度進行文字探勘，探討網路將客戶需求、軟硬體設備等資訊進行滿意度分析並發現其關聯性（Berezina, Bilgihan, Cobanoglu, & Okumus, 2016）；另一研究則利用社交媒體Facebook和Twitter上的非結構化文字內容，分析三個最大的比薩供應商的社群資訊，從而協助披薩業者分析競爭對手的情報（He, Zha, & Li, 2013）；陳裕菘、謝邦昌、李勝輝與陳郁婷（2014）將財經新聞網新聞文件作為文字探勘之目標，探討人民幣兌台幣之匯率變動相關新聞，透過文字探勘發現在短期預測上預測值、人民幣兌新台幣匯率與加權股價指數前三期有重要關聯性。文字探勘也可用於顧客關係管理，瞭解顧客消費行為與偏好，使企業發展新的行銷通路，找到適合不同顧客群之最佳行銷管道（朱瑀

馨，2007）；觀光旅遊方面，透過分析遊客的描述經驗，找出遊客對旅遊地之感受、態度和觀光價值（陳怡廷、陳麗如、吳姿瑩，2016）；醫療保健的應用上，大數據可用來檢視腦中風常見的方式、症狀和用藥之關聯性（丁怡婷、劉志光，2010）。由此可見，適用大數據及文字資料探勘的領域不勝枚舉，性議題也屬其一。

Stephens-Davidowitz（2015）利用Google搜尋關鍵字的出現頻率，發現男女性之性焦慮往往和另一半徹底相反，當中存在性迷思，像男性總是擔心陰莖尺寸太小，搜尋引擎上查找「如何讓增大陰莖」的次數，遠超過如何換輪胎、彈吉他、做歐姆蛋等日常事務；而女性時常擔心男性的陰莖尺寸過大，關鍵字「疼痛」於Google搜尋中最常與「做愛時」共同查詢；搜索次數方面，女性搜索伴侶陰莖尺寸問題的次數，與男性自行搜尋次數之比例是1：170，顯示女性與男性對於陰莖尺寸之期待差異甚鉅。男性於Google次常搜尋的問題為「如何讓自己更持久？」可見男性心裡總是在擔心自己會不會太快高潮，反觀女性焦慮的是「為什麼另一半這麼慢高潮」。研究發現Google搜尋中，「如何讓男友更快高潮」的次數幾乎和「如何更持久」一樣多；女性最常關注的議題是擔心私處有異味，尋求陰道異味方面的資訊。然而，目前以大數據資料分析性議題的研究相對匱乏，仍屬起步階段，本研究期望運用文字探勘的技術來探討現今大眾所關注的性議題及相關訊息，豐富性學研究之廣度。

三、批踢踢歷久不衰

批踢踢實業坊網路平台（簡稱PTT）是台灣最大的電子布告欄系統（bulletin board system, BBS），於1995年9月14日創立，目前由國立台灣大學電子布告欄系統研究社管理，以非商業性及學術性質為目的，提供網路上言論空間、電子郵件及聊天室等服務，盛行於大專院校中，為年輕族群最常使用的網路傳播媒介之一，可以經由手機、電腦、平板電腦等工具使用網際網路瀏覽；至今註冊總人數約150萬人，每日發表超過2萬篇新文章及50萬則推文，針對不同主題進行發文與討論。實務上，PTT有明確的板規、發文限制與罰則，使用者在板上的每篇發文，須經過板主及板工的審核，也會比照「台灣學術網路使用規範」 與「台灣學術網路BBS站管理使用公約」作為處分依據；違規的發文在通知後，若未在兩小時內改善或更正，板主將進行刪文動作，累積違規達五次之用戶，直接判定永久禁止發文。由此可知，PTT無論在使用者帳號或針對發文及回文的內容規範方面都是相當嚴謹的，因此其文字內容具代表性且潛在內容效度高，適合作為分析文本。

PTT的設置超過六千多個「看板」，分類按照不同主題內容及訴求，例如台灣大學、政治大學、大專院校（其他大學及專科）、青蘋果樹（各區的高中、國中、國小及補習班）、活動中心（社團、聚會、團體）、視聽劇場（偶像、音樂、廣電）、戰略高手（遊戲、數位、程設）、卡漫夢工廠（卡通、漫畫、動畫）、生活娛樂館（生活、娛樂、心情）、國家研究院（政治、文學、學術）、國家體育場等，其中熱門的看板有Gossiping版（八卦板）、sex板（西斯板）、feminine_sex板（女性西斯板）、StupidClown板（笨板）、Food板（美食板）等。討論態度以理性、尊重為第一條件，以知識性為輔。當中feminine_sex板所發表的話題非常廣泛，大多以女性經驗為中心，如孕事、性行為、性技巧、性經驗、生理反應、醫療經驗、新聞、性騷擾及性侵害等性事與女性事務，符合本研究所探究之主題。

龐大數量之文本中，舉凡外在生理結構、內在情慾幻想，都在本研究討論範圍內，前者期望協助女性正視自身身體現狀及變化，從中揣摩出與身體共處之道，並為自己的身體負責；後者則期望女性透過表達與抒發，進而正視自己的情慾世界。為了對不同面向的性議題進行系統化分類，本研究在性議題之分類也沿用美國性資訊與教育協會（Sex Information and Education Council of the United States, SIECUS）之全面性教育指南，其架構包含六大概念，包括（一）人類發展：涵蓋身體和生殖解剖、生殖、青春期、身體意象、性傾向與認同；（二）關係主題：涵蓋家庭、友誼、愛、約會、婚姻、承諾及養育；（三）個人技巧：涵蓋價值觀、作決定、溝通、決斷力、協商及求助；（四）性行為：涵蓋一生的性、自慰、性行為的分享、人類性反應、性幻想及性功能障礙；（五）性健康：避孕、墮胎、性傳染病與愛滋病毒感染、性虐待、生殖健康；（六）社會與文化主題：性與社會、性別角色、性與法律、性與宗教、性與藝術、性與媒體。

綜觀而言，人們溝通的習慣與網路密不可分，網際網路的盛行大開方便之門，供使用者分享資訊或尋求協助，而PTT具有匿名性及個人訊息的隱匿性，使用者會有較高意願分享與討論較為隱私的話題，其發文與回文是探討群眾對性議題之看法十分適合的資料來源，加上國內未曾運用大數據及文字探勘技術探討性相關議題，因此本研究以PTT中feminine_sex板的巨量資料為分析文本，運用文字探勘的技術與R語言（R language），進行資料蒐集與斷詞，然後分群、歸納與繪圖，最後分析大眾對於女性之性相關議題。

貳、研究方法

一、研究資料庫

本研究分析的資料為PTT中的feminine_sex板於2016年1月1日至2016年12月31日期間發表的文章，為期一年的時間，共1,438篇文章，利用網路爬蟲，抓取內文，形成語料庫，抓取的內容包含發文者ID、標題、時間、文章內容、回文等與性議題有關的發問及討論的文字資訊。

二、研究工具

本研究使用免費軟體R語言作為主要工具。根據國外知名數據挖掘網站KDnuggets，從2014年至2017年各年度全球各項資料探勘工具使用率的調查發現R語言皆占將近50%；於2016年KDnuggets調查，資料探勘軟體前5名工具使用調查，結果R語言是優於商業軟體Tableau、Microsoft、SQL Server、MATLAB及免費軟體Python，R語言相容性是最高的（Piatetsky-Shapiro, 2017）。R語言目前擁有一萬多個基本套件，為特定資料分析所設計的指令，當中包括系統預設或由核心團隊所釋出的指令組合，也有R語言的使用者與愛好者所貢獻的。以下四點分析R語言在資料探勘的優勢：

（一）功能性

R語言可自行於網路下載，並且開放原始碼軟體，提供個人化修改過程中重要的功能性。R語言是數據採礦過程的核心組件，能將結構及非結構化的資料有效整合。

（二）強大視覺化功能

R語言的繪圖套件相當豐富，包含一般利用函數繪製的統計圖表如繪圖莖葉圖、盒型圖、直方圖等；互動式圖表，將滑鼠遊標停留於圖形中，即可顯示資料訊息及數值，並能將圖表輸出到網頁上等功能；工業製程上高階繪圖的完整圖表，並自動產生座標、標籤及標題；還能依使用者需求，透過函數更改定義。由此可知，R語言具有強大的統計分析及視覺化功能，能使用PDF、JPEG、PNG等不同格式輸出圖片，支援共各種統計及圖形顯示，將資訊視覺化。

（三）套件發展性與不斷更新

R語言允許任何人對其進行修改及支援，使用人數越多，則對套件的貢獻也越多，截至2018年10月CRAN上發布高達13,285種套件，且持續增加中。其特點是能透過使用者撰寫的套件，來增強軟體的學習能力，網路上有來自世界各國的R語言使用者，經常會開發或更新不同套件以擴充功能，其他使用者也可依照自己的需求下載與更新。

（四）相容性及擴充性強

R語言的原始碼可自由下載使用，且高相容性能應用於Windows、Linux、IOS等不同作業系統，及32或64位元的處理器上執行。此外，R語言能從Microsoft SQL、Excel、Stata、Minitab、MATLAB、SPSS、SAS等軟體，匯入其他不同格式的資料、讀取資料集及完美結合各式工具。

三、資料儲存

本研究使用R語言透過撰寫的爬蟲程式碼於PTT進行網路爬蟲，抓取feminine_sex板一年度之性相關資料，探勘前置處理、關鍵字提取、詞項文件矩陣、TF-IDF（term frequency-inverse document frequency）、及資料分析等五部分。首先文字探勘之前置處理，包含偵錯（anomaly detection）、轉換文字（transforming text）、檢視文字（retrieving text）、中文斷詞（Chinese segmentation）、剔除非必要之元素。透過網路爬蟲，抓下PTT網頁內容，除了內文與回文外，尚包含文章發布的標題、作者、時間以及推、噓等回文的標記。因此爬取文章後，須先經過自然語言處理（natural language processing, NLP），方能進行接續步驟。

（一）偵錯

提取文章前經由程式碼判斷後刪除空值，留下正確的文章，以避免擷取過程中網頁格式錯誤或文章刪除，所產生的網頁錯誤的問題。

（二）轉換文字

將網路文章內容轉換為文檔形成語料庫，將語料庫編碼轉為程式讀得懂的數值。

（三）檢視文字

本研究使用PTT之feminine_sex板文本來進行大數據挖掘，提取文章內容，並使用程式碼檢視文本內容。

（四）中文斷詞

透過中文斷詞套件，對於文章字資料進行斷句。對於文字資料而言，每一個「字詞」皆是代表語義的最小單元，因此在進行自然語言處理時，對文章中的文字做斷詞的動作。

（五）剔除非必要之元素

係由於發文作者、時間、板規及其他特殊標記等會對分析結果產生不必要影響及錯誤標籤，故於資料前置處理後，本研究刪除「編輯、標題、時間、發信、實業、作者、文章、網址、東西、照片、時候」等主題無關的詞彙，並移除英文、數字、符號及停用字，最後只保留名詞。

四、資料流程

歸納國內外相關文獻對於文字探勘系統架構之介紹，本研究將文字探勘之流程系統分為下列三項要件：

（一）文字矩陣

第一部分為詞項文件矩陣，透過套件建立詞項及文件的矩陣（term-document marix, TDM），每一欄表示一篇文章，而每一列表示一個字詞，矩陣中的數字則代表每一個字詞在各篇文章中出現的次數，也能藉由交換行列來建構文檔詞彙矩陣（document-term matrix）。

（二）TF-IDF

第二部分為TF-IDF，TF-IDF演算法包含了兩個部分，詞頻（term frequency, TF）跟逆向文件頻率（inverse document frequency, IDF），是一種加權技術及統計方法，常用於評估該詞彙對於文件集或語料庫的重要程度。TF係指某一個給定的詞語在該文件中出現的頻率，第 t 個詞出現在第 d 篇文章的頻率記作$tf_{t,d}$ IDF詞彙 t 總共在 d_t 篇文章中出現過，則詞彙 t 的IDF

定義成 $idf_t = \log\left(\frac{D}{d_t}\right)$。如果詞彙 t 在非常多篇文章中都出現過，就代表 d_t 很大，此時 idf_t 就會比較小。TF-IDF值即是TF值與IDF值之乘 $tfidf(t, d, D) = tf(t, d) \times idf(t, d)$是以頻率而不是次數來看待文字的重要性，讓文章與文章之間比較有可比較性。

（三）狄利克雷分配（latent dirichlet allocation, LDA）

第三部分為LDA，為一種主題模型，利用一組連續的多變量機率分布，將語料庫中每篇文檔的主題按照機率分布的形式呈現。它是一種無監督學習算法，需要的僅僅是語料庫以及指定主題的數量k值即可，LDA對於每一個主題均可找出詞語來描述，目前在文本挖掘領域包括文本主題識別、文本分類以及文本相似度計算方面都廣泛應用。

歸納國內外專家學者對於文字探勘的功能之研究，大致可分為：文字雲（word cloud）、進行關鍵詞的觀察（frequent terms and associations）、群集分析（cluster analysis）、網絡分析（network analysis）、主題分析（topic model）、關聯分析（associative analysis）等六步驟，茲就析述如下。

1.文字雲

構建一個矩陣之後，視覺化圖片顯示出詞彙的重要性，即出現頻率較高的字詞於文字雲中的尺寸也越大。

2.關鍵詞彙

檢視詞頻較高之詞彙。係使用統計學角度來衡量字詞的重要性，檢視該字詞在同一篇文章或語料庫中的出現次數，藉此瞭解詞彙之重要性及文章之特性，故從大量非結構化的文字中找出關鍵字為本研究相當重要的目標之一。

3.群集分析

係依照字詞與字詞之間的距離及相似程度進行分類，目的是將相似的詞彙歸類。利用相似度或相異度將字詞分群歸屬到數個群集，同一群集內的字詞相似程度較大，不同群間則相似程度小。透過集群分析，能找出哪些字詞歸於同一類，就分群結果的特徵及意涵加以探討。

4.網絡分析

網絡圖係用來表示系統，並由節點和鏈組成。節點是被鏈連接，而鏈表示節點間的關係，全部的節點和鏈集合為網絡圖。係提供使用者從大量的文本中快速瞭解內容重點，並透過詞彙的權重瞭解關鍵詞彙之間的相關性。每一篇文章順著一起事件的網絡發展，有時候可能發生不同網絡但有相同節點重疊的情況，表示可能發生一起交叉的事件，而重疊的詞彙往往都是語料庫中出現頻率較高的核心字詞。

5.主題分析

本研究採用主題模型並使用LDA對文章進行潛在主題分析，透過相關詞彙，找出文件中潛在的主題變量，目的在於從文本中發現隱藏的主題，找出大眾所關注的性議題，並對各個主題進行命名分析。

6.關聯分析

構建術語文檔矩陣，以表示術語和文檔之間的關係，其中直行代表詞彙，橫列代表文檔，目的為使用函數檢視語料庫的詞彙並進行分析。關聯分析的目的在於產生部分資料的概要，例如尋找資料子集間的連結關係或資訊與資訊間衍生的關係。Cabena、Hadjinian、Stadler、Verhees與Zanasi（1998）定義三種鏈結分析的模式：Association discovery、sequential pattern discovery與Similar time sequence discovery。Association discovery可用於顧客購買的商品分析，此方式亦稱為菜籃分析（market basket analysis ）；sequential pattern discovery是藉由顧客過往的購買交易來分析客戶長期的購買行為；similar time sequence discovery是用來分析商品銷售曲線與銷售時間點的關聯性。

參、研究結果與討論

本研究透過R語言網路爬蟲自動蒐集feminine_sex板一年份文章共1,438篇，一篇文章各存成一個文檔以形成語料庫，運用文字探勘技術之文字雲、關鍵詞彙、集群分析、網絡分析、主題模型、詞彙關聯等方式，分別進行演算及分析，探討社會大眾關注之性議題。

一、文字雲

本研究分析文本為PTT之feminine_sex板文章，建置語料庫後以自然語言處理、中文斷詞，產生184,145個詞彙，然後進行名詞、英文、數字、符號及停用字（stop words）之篩選，留下4,934個詞彙，在操作指令設定min.freq = 120，將出現次數超過120次的詞彙加以統計，結果得到120個詞彙。將這些較常出現在feminine_sex板的詞彙可視化可得出文字雲（圖1），圖中詞彙的圖像越大，表示該詞彙出現頻率越高，亦可推論該詞彙於feminine_sex板是相對重要之議題，依序為醫生、問題、男友、感覺、女性、女生、月經、女友、身體、經驗、建議、陰道、醫師、子宮、關係、婦產科等詞彙。

二、關鍵詞彙

資料視覺化是文字探勘中，呈現資料的一個重要方法，使用R語言之ggplot2套件，將feminine_sex板語料庫中出現次數前20的詞彙，以長條圖的方式呈現，橫軸為出現次數，縱軸則為詞彙，能清楚比較不同詞彙間出現頻率的差異。feminine_sex板語料庫經由函數進行統計後，形成feminine_sex板詞頻長條圖（圖2），前三高分別為醫生（2,568次）、問題（2,430次）、男友（2,201次），其後依序為感覺（1,900次）、女性（1,815次）、女生（1,206

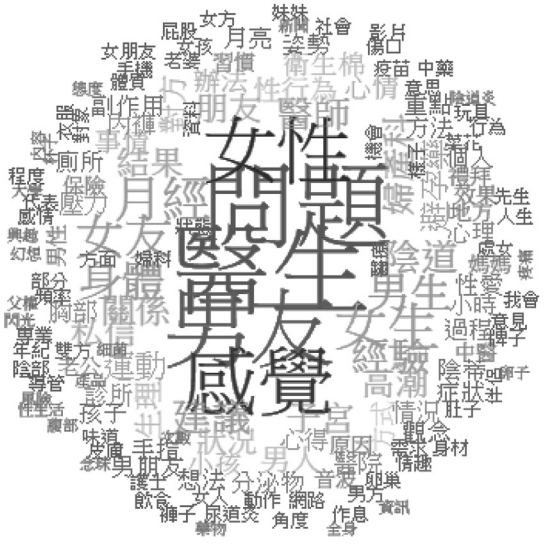

圖1　feminine_sex板文字雲

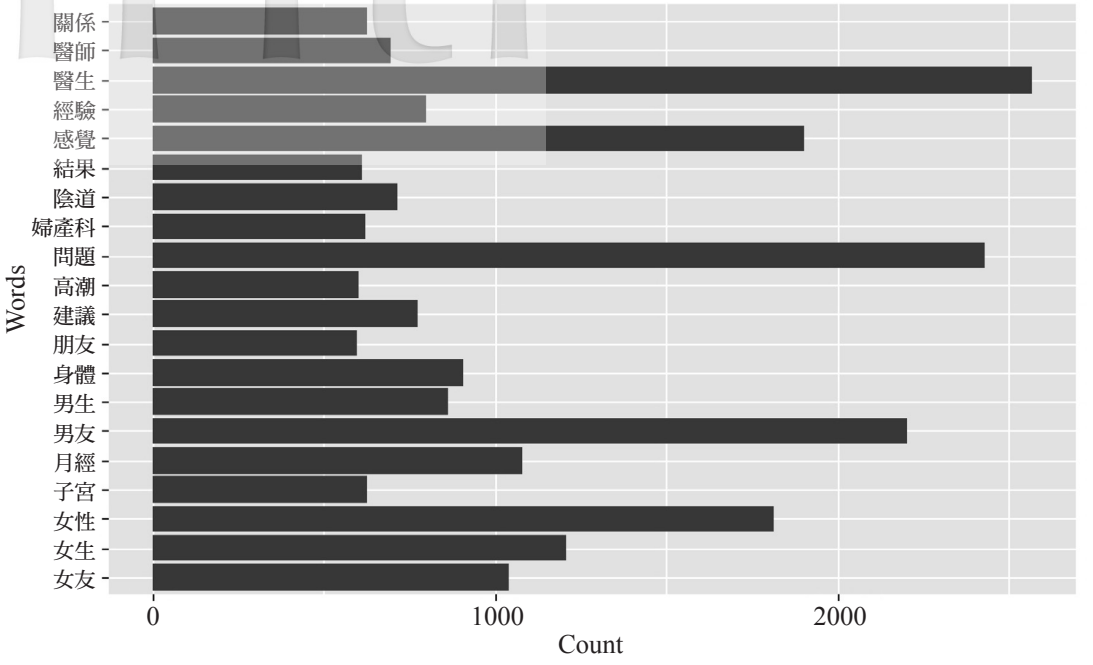

圖2 feminine_sex板詞頻長條圖（縱軸依筆畫順序排列）

次）、月經（1,082次）、女友（1,041次）、身體（906次）、男生（862次）、經驗（797次）、建議（775次）、陰道（714次）、醫師（694次）、子宮（628次）、關係（627次）、婦產科（621次）、結果（610次）、高潮（600次）、朋友（597次）。由此推論醫療衛生相關問題是相當受女性族群關注之議題，尤其高達一半以上之詞彙皆與醫療衛生有所關連，其討論脈絡屬於美國性行為資訊與教育委員會（Sexuality Information and Education Council of the United States, SIECUS）六大概念中「性健康」及「人類發展」。至於男友、女友、關係、朋友都是人際關係的詞彙，涵蓋了SIECUS六大概念中「關係」之下友誼、約會及愛的領域。

三、集群分析

集群分析是依照字詞之間的距離及相似程度進行分類，透過詞彙TF-IDF進行稀疏矩陣處理，使同一群內的字詞相似程度大，各群間的相似程度小。透過集群分析，相似度及相異度將詞彙分群歸屬到數個群集，能夠看出哪些詞彙會被歸在同一類，並對分群結果的特徵及其所代表的意義加以解釋及探討。透過R語言套件函數ward.D（詞彙的最小變異法），產生樹狀架構之feminine_sex板詞彙階層圖（圖3）。

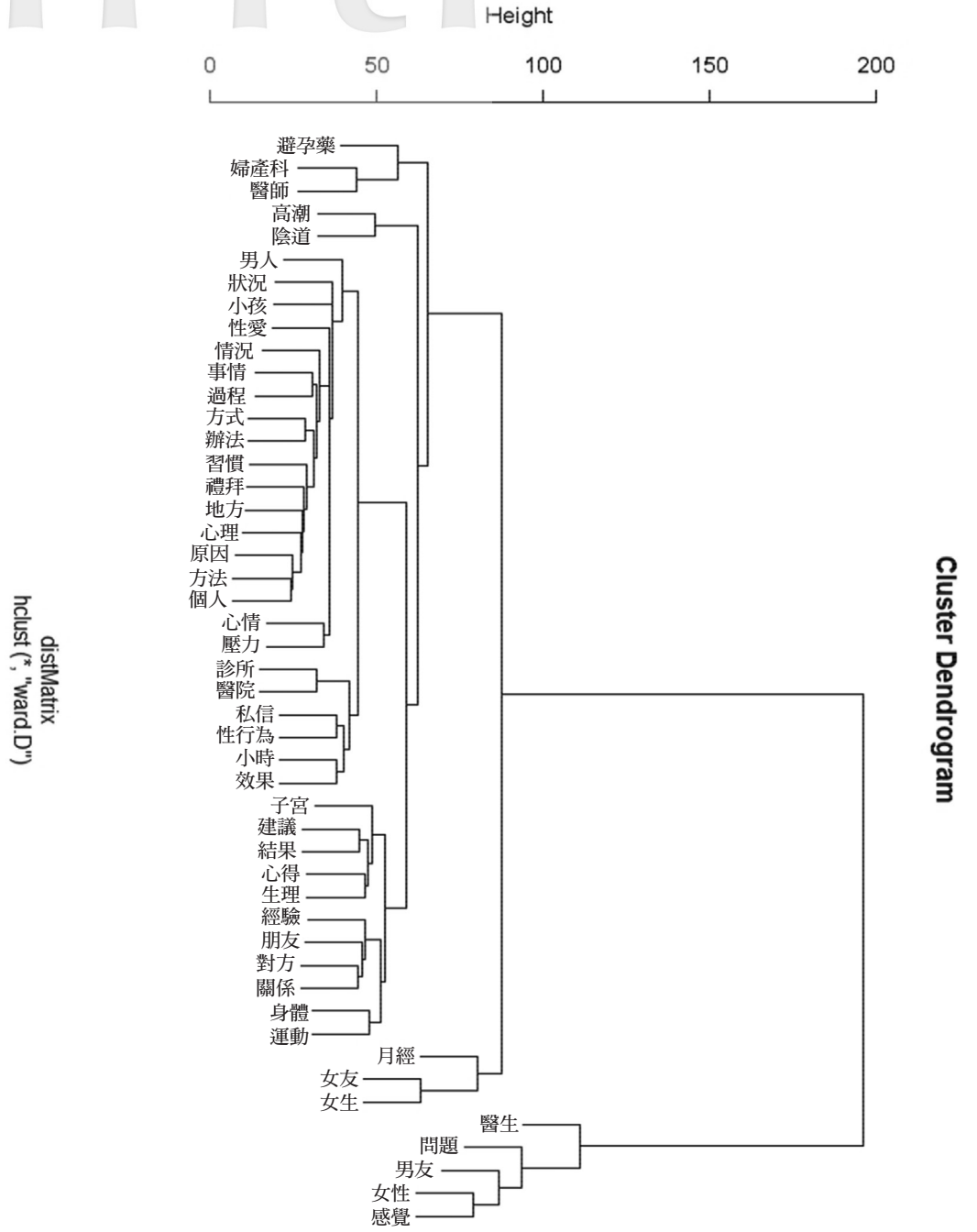

圖3　feminine_sex板詞彙階層圖

feminine_sex板詞彙階層圖（圖3）第一階層以醫生、問題、男友等詞彙為主，推論女性對於性健康議題十分關注，並且會於網路尋求解答與討論，而男朋友的話題討論也是焦點之一。第二階層以女生、女性及月經，推論是關於生理期護理保健之性健康之議題。第三階層討論的有醫師、避孕藥、婦產科，也是性健康相關議題。其後分支有兩組包含診所、醫院、時效、子宮、生理、身體、運動，可推測與性健康也是息息相關，另外，有分支討論高潮與陰道，是屬於性行為之範疇。綜上所述，集群階層分析的結果發現feminine_sex板使用者對於性健康議題的討論涵蓋了各階層，而部分段階層討論到陰道及高潮有關之性行為議題。

四、網絡分析

網絡圖著重於統計數字的意義，而非可視化圖像，若兩個詞彙同時出現在同一篇文章中，則其權重 + 1。feminine_sex板權重值（表一）處理後形成feminine_sex板網絡圖（圖4），共有12個節點及66條連結線，其中關聯性最大之詞彙為「女性與問題」權重值為597，依序權重值：「女性與私信」為574、「醫生與問題」為527、「感覺與問題」為538、「女性與感覺」為465、「私信與問題」為444、「男友與問題」為404、「男友與女性」為382、「建議與問題」為366、「經驗與問題」為364。

表一　feminine_sex權重值

變項	1	2	3	4	5	6	7	8	9	10	11	12
1. 女生	—	310	260	200	174	155	346	150	299	195	136	191
2. 女性		—	382	574	269	264	597	207	465	295	292	245
3. 男友			—	299	174	175	404	155	342	194	184	211
4. 私信				—	186	185	444	139	339	206	230	169
5. 身體					—	161	315	146	260	152	206	152
6. 建議						—	366	151	237	171	260	135
7. 問題							—	303	538	364	527	306
8. 結果								—	224	150	202	143
9. 感覺									—	279	288	239
10. 經驗										—	200	168
11. 醫生											—	160
12. 關係												—

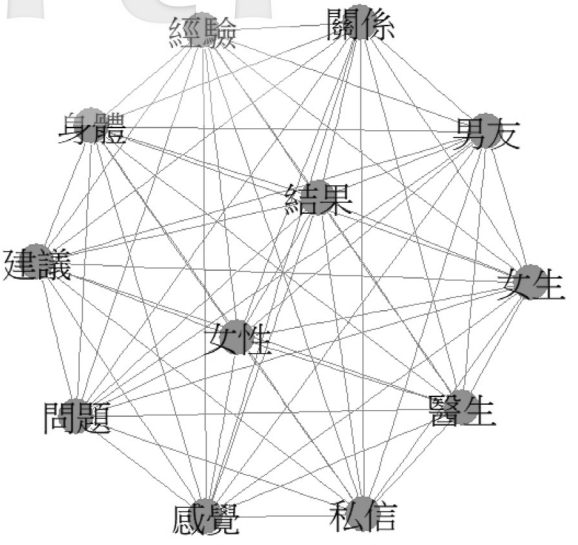

圖4　feminine_sex板網絡圖

從feminine_sex板網絡圖（圖4）與權重值（表一），發現「問題」與其他詞彙之關聯性很高，以此可推論feminine_sex板的使用者大多是提出個人問題，希望尋求協助與討論。查找「女性與問題」兩詞關聯性較高的文章，其中有關於女性詢問大眾是否會因為性功能障礙的問題而分手及女性生理之相關問題。「女性與私信」相關文章中，有關於女性衛生醫療用品之使用經驗或詢問，並希望網友能夠私信告知。「感覺與問題」相關文章裡面，有關女性性行為時沒有感覺之問題及性生活不協調之問題等文章。「女性與感覺」相關文章中，有關於女性無法藉由性愛得到性滿足，及女性性行為時陰道沒有感覺等問題。「私信與問題」相關的文章發現大多詢問藥物的問題。「問題與男友」相關文章裡，有關於男友不願意戴保險套、男友性交避孕及男友在性行為的動作與偏好之相關問題。「男友與女性」相關文章，大多是以女性角度談論男友會對哪些類型的女性有興趣，及男友與其他女性間的關係。「建議與問題」詞彙相關的文章大多在談論女性性健康相關問題。「經驗與問題」相關文章，有關於大眾分享性經驗與性相關疑問。「女生與問題」相關文章，多為談論性行為及性關係上的問題，像是女生是不是較被動及親密關係溝通的問題。

五、主題模型

TF-IDF衡量了各詞彙對於文本的重要性，以便於利用主題模型對文本進

行分類，經過R語言NbClust及Cluster套件決定最佳分群指標藉以評估分群的效果。本研究將最小的分群數設定為2，最大分群數設定為8，使用K-Means集群分析演算法執行分群運算，得到feminine_sex板主題模型分群指標（圖5），分析結果顯示13個指標認為語料庫最佳群集數為3群，5個指標認為最佳群集數為4群，其餘為2群、6群、8群。feminine_sex板NbClust套件分析圖（圖6）之中，折線圖顯示集群數目增加則複雜度會下降，從2群至3群快速下降，3群後趨緩。綜合K-Means集群分析演算法及NbClust套件分析圖，feminine_sex板資料分為3個集群數為最佳。

六、cluster分群套件

R語言cluster套件在clusGa函數中，可以用來評估資料的最佳集群數，cluster在本研究使用kmeans演算法，將觀察值分類為不同集群，計算出觀察值的最佳數量的集群數，先設定在1到10之間，為了能得到較精確的結果，將B值設定在150，統計分析解果如圖7及圖8，在分為4群時複雜度上升，模型的解釋例會下降，經統計結果顯示當K值為3是最佳集群數。經由NbClust及cluster兩個套件的驗證結果皆顯示，該語料庫的主題模型，K值為3是最佳集群數。

主題模型係將NbClust之K-Means集群演算法的驗證結果，經Topic Model

```
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 13 proposed 3 as the best number of clusters
* 5 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 2 proposed 8 as the best number of clusters


***** Conclusion *****


* According to the majority rule, the best number of clusters is  3
```

圖5　feminine_sex板主題模型NbClust套件分群指標

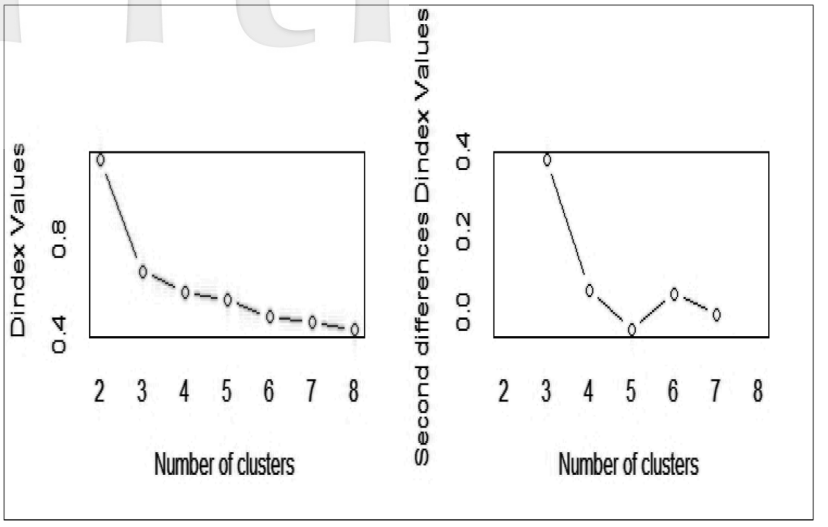

圖6　feminine_sex板NbClust套件分析圖

```
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = dtm4, FUNcluster = kmeans, K.max = 10, B = 150, v
erbose = interactive())
B=150 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 3
          logW     E.logW          gap          SE.sim
 [1,] 8.608395  10.41280  1.804403  0.002889244
 [2,] 8.558525  10.37330  1.814773  0.003089302
 [3,] 8.531507  10.34783  1.816323  0.003089619
 [4,] 8.508324  10.32728  1.818954  0.003101406
 [5,] 8.505861  10.31262  1.806757  0.002745204
 [6,] 8.487084  10.30033  1.813242  0.002662800
 [7,] 8.468878  10.28932  1.820444  0.002579209
 [8,] 8.476117  10.27926  1.803147  0.002752000
 [9,] 8.456613  10.27195  1.815337  0.002684581
[10,] 8.442026  10.26513  1.823104  0.002621180
```

圖7　cluster套件分析結果

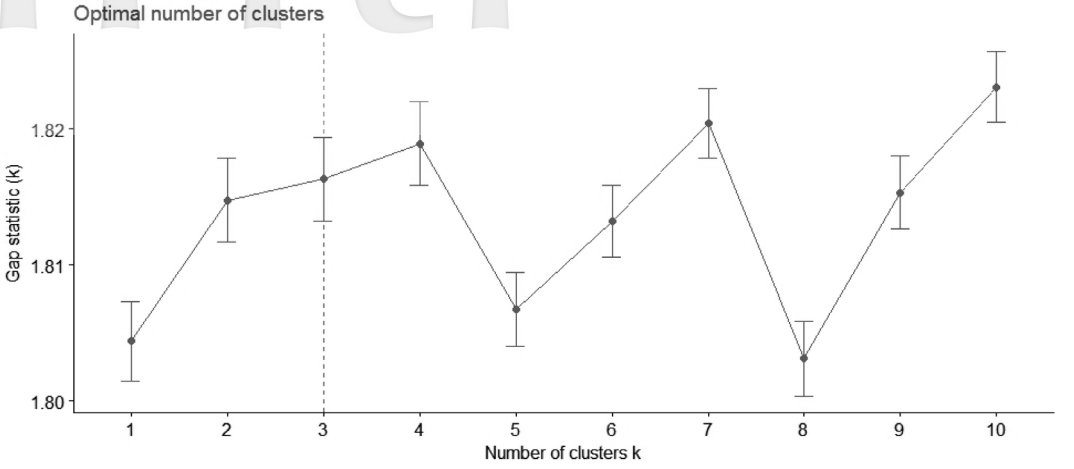

圖8 cluster套件分析圖

套件中潛在LDA函數，依最佳集群數將feminine_sex板K值設定為3，使用terms()函數檢視各主題30個關聯詞彙（表二），依其詞彙，分別將該主題命名為親密關係、避孕諮詢、衛生醫療。

本研究焦點之一是從大數據資訊中，利用演算法分群，從各式的性議題尋找出潛在重要主題，形成主題模型，瞭解大眾的性心理及對性議題之討論脈絡。首先feminine_sex板主題模型共分為3組，分別命名為親密關係、避孕諮詢及衛生醫療。主題一為親密關係，包括男友、感覺、女性、女生、問題、女友、高潮、對方、男人、經驗、關係、身體、陰道、私信等字詞，相關詞彙主要是探討伴侶關係中的心情感受及問題，或性行為、性健康、性經驗等相關議題。主題二為避孕諮詢，包括避孕藥、月經、問題、小孩、保險套、媽媽、副作用、孩子、效果、身體、女友、小時、建議、女性等字詞，主要是探討生理性健康，及詢問避孕藥廠牌及分享使用經驗，或討論保險套的使用經驗及感受分享等相關議題討論。主題三為衛生醫療，包括醫生、問題、醫師、婦產科、子宮、月經、感覺、衛生棉、女性、建議、診所、醫院、陰道、分泌物等字詞，文章大多討論女性婦科方面的疾病及問題，許多文章為推薦診所、醫生或看診後的經驗分享。

女性之性心理及性議題，主要分為親密關係、避孕諮詢及衛生醫療三個脈絡，親密關係部分涵蓋性行為及性健康相關問題的討論，而避孕諮詢及衛生醫療皆為性健康之相關範疇，由此可知，面對性相關議題時，女性心理十分重視生理健康，其次才是性行為的相關問題與感受。

表二　feminine_sex板主題命名及詞彙

主題一 親密關係	主題二 避孕諮詢	主題三 衛生醫療
男友	避孕藥	醫生
感覺	月經	問題
女性	問題	醫師
女生	小孩	婦產科
問題	保險套	子宮
女友	媽媽	月經
高潮	副作用	感覺
對方	孩子	衛生棉
男人	效果	女性
經驗	身體	建議
關係	女友	診所
身體	小時	醫院
陰道	建議	陰道
私信	女性	分泌物
朋友	牌子	內褲
運動	性行為	生理
性愛	胸部	狀況
事情	結果	身體
心理	皮膚	中醫
心情	觀念	經驗
陰蒂	朋友	結果
辦法	方式	音波
男性	感覺	運動
建議	女孩	心得
老公	事情	廁所
胸部	辦法	症狀
姿勢	劑量	尿道炎
方式	關係	護士
壓力	男友	醫療
手指	風險	私信

七、詞彙關聯分析

詞彙關聯分析是選擇語料庫中的高頻詞彙，找出與其關聯性較大的相關字詞，特徵為挖掘詞彙產生的規則，用來描述詞彙間的關聯性，亦表示該二字詞較常同時出現於一篇文章之中。本研究詞彙關聯分析部分使用findAssocs指令，分析feminine_sex板頻率高的詞彙。

與「醫生」相關聯的字詞中，關聯性最高為「醫師」相關性0.49，依序是婦產科（0.46）、超音波（0.44）、子宮（0.39）、婦科（0.39）、態度（0.36）、診所（0.35）、建議（0.32）、醫院（0.31）、狀況（0.3），推論女性對子宮、婦科方面十分重視，並且會分享就診某醫生或醫療院所之就醫經驗，屬性健康之範疇。

「男友」一詞關聯性最高的字詞為「男生」，相關性為0.35，依序為感覺（0.32）、女生（0.29）、信心（0.29）、男人（0.27）、胸部（0.27）、男方（0.25）、頻率（0.25）、關係（0.25）、視覺（0.23）。對於女性而言，男友、男性是性相關議題中重要主題，且在乎感覺、關係，女性的性心理對於身材、胸部會有所焦慮，查找中發現女性對於身體意象的關注，並且也會在意男友對於胸部的偏好、性吸引力及滿意度。

「高潮」一詞也是熱門的話題字詞，與其關聯性最高的字詞為「陰蒂」相關性0.6，依序為陰道（0.33）、感覺（0.27）、姿勢（0.22）、成就感（0.21）、肉體（0.21）、模式（0.21）、玩具（0.2）、陰莖（0.2）、男友（0.17）。討論的內容有關於陰道高潮、陰蒂高潮、透過玩具達到高潮、性交過程中嘗試了各種姿勢仍然無法達到高潮，及高潮相關感受。

最後的討論焦點是「避孕藥」，關聯性最高的字詞為「月經」，相關係數為0.34，依序是副作用（0.33）、藥物（0.28）、原理（0.26）、新藥（0.25）、指數（0.24）、藥效（0.23）、卵巢（0.22）、規律（0.22）、保險套（0.21）。女性對於性健康相關議題十分關注，尤其是避孕藥物之副作用、藥效，也會討論到保險套，推論女性對於避孕方式的選擇十分重視，而避孕藥是熱門的選擇及討論項目，透過網路獲得及分享該藥物之特性、原理、副作用、藥效等知識，並且相互交流。

肆、研究結論

本研究之樣本以2016年1月至2016年12月間PTT論壇feminine_sex板所發表的文章，運用文字探勘技術，將非結構化之文字資訊進行結構化處理，並以非監督式學習演算法的概念，剖析大數據之文字訊息，進行大眾對性議題的關注。相較於內容分析，文字探勘於文本的複雜度、件數與分析之詞彙數明顯有著數量及技術的優勢。關於性議題大眾已經習慣透過文字訊息在網路上閱讀、分享和交換資訊，並已成為一種趨勢，網路上關於性問題的資訊不單單只有看它如何形成及如何去對應，換個方式也可以發現文字有凸顯出問題的效果，詞頻亦能反映大眾所關注的議題，經由討論的主題與相關詞彙，在性議題上無論是正面或負面的反饋，也能表露出網路使用者的性行為、性知識以及對性問題隱含的樣貌，本研究透過研究結果對未來的研究方向提出建議。

綜合研究結果，語料庫經中文斷詞後產生28,129個詞彙，進行篩選名詞及去除停用字後留下4,934個詞彙，feminine_sex板出現次數最頻繁的三個詞彙依次數排序為醫生、問題、男友。依據主題模型我們從資訊中找出潛在類別規則語料庫關鍵字與主題之間的關聯再做命名對照，分析結果呈現潛在的主題大多圍繞在親密關係、避孕諮詢以及衛生醫療等三個主要議題。可見使用者最關注性健康，以避孕、醫療資訊相關訊息的討論為大宗。其中避孕議題也呼應了Merzel等人（2004）的研究觀點，性健康話題的總體方面沒有性別差異，但女性比男性更可能討論避孕。次而論及男女性行為相關問題，如女性高潮、身體意象及性吸引力，表示親密關係也是大眾另一關注焦點，這與Google搜尋引擎的大數據分析（Stephens-Davidowitz, 2015）不謀而合。

依SIECUS性教育指南的六大概念分類，民眾關注的性議題主軸在於「關係」、「性行為」與「性健康」這三個主要概念，此研究結果亦可提供教育及醫療衛教相關單位參考，有助掌握大眾討論性議題的模式與習慣，從他們的網路活動瞭解對性教育的實際需求，補強前線工作者實施性教育的成效，並提供相關研究的新方向。

參考文獻

丁怡婷、劉志光（2010）。文字探勘技術應用於中醫診斷腦中風之研究。**數據分析**，**5**(4)，41-64。

古鐘响（2009）。**黃色笑話收集與性學分析研究**。未出版之碩士論文，樹德科技大學人類性學研究所，高雄市。

朱瑀馨（2007）。**運用資料探勘技術於人壽保險業顧客關係管理之研究**。未出版之碩士論文，淡江大學保險學系保險經營研究所，台北縣。

黃文、王正林（2015）。**利用R語言打通大數據的經脈**。台北市：佳魁資訊。

陳怡廷、陳麗如、吳姿瑩（2016）。從部落格探索客家旅遊目的地意象之研究——自然語言處理的方法與應用。**戶外遊憩研究**，**29**(2)，81-111。

陳裕菘、謝邦昌、李勝輝、陳郁婷（2014）。運用文字探勘與資料採礦技術建立匯率預測模型——以人民幣兌新台幣為例。**數據分析**，**9**(1)，133-146。

曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯（2005）。**資料探勘（Data Mining）**。台北市：旗標。

鄭天澤、陳麗霞、楊亨利、胡正文、鄭閔安（2017）。**2017年台灣寬頻網路使用調查報告**。台北市：財團法人台灣網路資訊中心。

鄭天澤、楊亨利、陳麗霞、胡正文、劉千鳳（2015）。**2015年台灣寬頻網路使用調查報告**。台北市：財團法人台灣網路資訊中心。

Adriaans, P., & Zantinge, D. (1996). *Data mining*. Harlow, UK: Addison Wesley.

Aggarwal, C. C. (2015). *Data mining: The textbook*. Cham, Switzerland: Springer International.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, *25*, 1-24.

Berry, M. J. A., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. New York: John Wiley & Sons.

Blake, C. (2011). Text mining. *Annual Review of Information Science and Technology*, *45*(1), 121-155.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Upper Saddle River, NJ: Prentice-Hall.

Cooper, A., Delmonico, D. L., & Burg, R. (2000). Cybersex users, abusers, and compulsives: New findings and implications. *Sexual Addiction & Compulsivity: The Journal of Treatment & Prevention*, *7*, 5-29.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*, 1707-1720.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, *57*, 321-326.

Han, J., & Kamber, M. (2001). Data mining: Concepts and technologies. *Data Mining Concepts Models Methods & Algorithms*, *5*(4), 1-18.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*, 464-472.

Merzel, C. R., Vandevanter, N. L., Middlestadt, S., Bleakley, A., Ledsky, R., & Messeri, P. A. (2004). Attitudinal and contextual factors associated with discussion of sexual issues during adolescent health visits. *Journal of Adolescent Health*, *35*, 108-115.

Mishra, P. (2016). *R data mining blueprints*. Birmingham, UK: Packt.

Moreira, E. D. Jr., Brock, G., Glasser, D. B., Nicolosi, A., Laumann, E. O., Paik, A., et al. (2005). Help-seeking behaviour for sexual problems: The global study of sexual attitudes and behaviors. *International Journal of Clinical Practice*, *59*, 6-16.

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing & Management*, *42*, 785-804.

Piatetsky-Shapiro, G. (2017). *Python overtakes R, becomes the leader in data Science, machine learning platforms*. Retrieved 6 30, 2018, from https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Plaud, J. J., Gaither, G. A., & Weller, L. A. (1998). Gender differences in the sexual rating of words. *Journal of Sex & Marital Therapy*, *24*, 13-19.

Sanders, J. S. (1978). Male and female vocabularies for communicating with a sexual partner. *Journal of Sex Education and Therapy*, *4*, 15-19.

Stephens-Davidowitz, S. (2015). *Searching for sex*. Retrieved 6 30, 2018, from https://www.nytimes.com/2015/01/25/opinion/sunday/seth-stephens-davidowitz-searching-for-sex.html

Sullivan, D. (2001). *Document warehousing and text mining: Techniques for improving business operations, marketing, and sales*. New York: Wiley.