Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

《傳播與社會學刊》，（總）第59期（2022）：177–209

專輯論文

基於認知的微博用戶聚類方法研究——以「湯蘭蘭事件」為例

周勝

摘要

為解決微博用户屬性數據和評論文本數據較少甚至缺失情況下的用户聚類問題，本文提出了一種基於用户認知差異的微博用户聚類方法：根據用户在關注和轉發資訊源上的判斷和選擇，構建用户和資訊源的雙模網絡（two-mode network），通過雙模網絡中二部圖（bipartite graph）的切割，實現了用户的聚類。採用譜聚類（spectral clustering）方法，建立混合的關注和轉發兩級聚類處理方式，能夠有效區分不同認知屬性的用户群體，精準觀察資訊在社群中的傳播機制。以中國互聯網上的一個熱點事件「湯蘭蘭事件」為例，介紹微博用户聚類方法的操作流程和評估標準。實際聚類結果表明：該方法在劃分群體規模和準確性上，綜合性能較好，對於社交平台用户的群體劃分、行為預測和輿情分析，具有較強的實際應用價值。

關鍵詞：微博用户、雙模網絡、聚類

周勝，武漢學院藝術與傳媒學院傳播系講師。研究興趣：傳播社會學、大數據、計算傳播。電郵：zhousheng78@163.com

論文投稿日期：2021年2月10日。論文接受日期：2021年7月26日。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

Communication and Society, *59* (2022), 177–209

Special Issue Articles

Research on Microblog User Clustering from a Cognitive Perspective

Sheng ZHOU

Abstract

Focusing on microblog user attributes and missing content data, a microblog users clustering method based on users' cognitive differences is proposed. By referencing users' selections and judgments regarding the information sources they follow and forward, we can identify the typical characteristics of group members; we can also use this information to analyze users' behaviors and attitudes while they are participating in various kinds of groups. On the basis of users' perceived dissimilarity regarding microblog topics, a two-mode network comprised of user and media is built. The clustering of users is then realized by cutting the bipartite graph in the network. By applying the spectral clustering method to the hybrid mechanism of two clustering stages (i.e., followed topics and forwarded topics), we can effectively distinguish between user groups with different cognitive attributes; this also enables us to accurately observe the transmission mechanism of information within the community. Taking the Tang Lanlan event as an example, this paper introduces specific operation and evaluation criteria of the microblog user clustering method. Upon thorough consideration of the scale and accuracy of the group division, the results show that the comprehensive property is

Sheng ZHOU (Lecturer). Communication Department, School of Art and Media, Wuhan College. Research interests: communication sociology, computational communication, big data.

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

improved with use of the proposed clustering method. The actual results demonstrate the precision and effectiveness of the proposed method for the purposes of group division, behavior prediction, and public opinion analysis of social platform users.

Keywords: microblog user, two-mode network, clustering

Citation of this article: Zhou, S. (2022). Research on microblog user clustering from a cognitive perspective. *Communication and Society, 59*, 177–209.

致謝

本研究受到香港中文大學「第十四屆傳播學訪問學者計劃」的支持，受2021年度湖北省高等學校哲學社會科學研究項目《基於認知和傳播行為的在線社交網絡用戶聚類方法研究》和武漢學院科研創新團隊（KYT202001）的支持。作者衷心感謝武漢學院鄧青教授以及《傳播與社會學刊》諸位匿名評審、編輯對本文提出的意見和建議。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

研究背景與研究問題

微博作為資訊驅動型的社交網絡平台，通過關注機制，分享即時的簡短資訊，具有傳播的高時效性和參與的開放性。截至2020年9月，微博日活躍用戶達到2.24億，月活躍用戶5.23億，成為擁有大量用戶群體、龐大信息量和廣泛社會影響力的媒體平台（新浪微博數據中心，2021）。

用戶的聚類，是線上社交網絡分析的一個基本任務。將數量巨大的微博用戶，按照特徵差異和相似度進行聚類和群體劃分，是實現社交網絡的結構分析、用戶屬性和行為推測、用戶推薦和輿情監測等眾多功能的前提和技術實現途徑，為線上網絡社區發現、用戶行銷、輿情態勢分析、資訊源的定位等實際問題，提供分析工具。商業應用層面：席運江、杜蝶蝶、廖曉及仉學紅(2020)的研究成果發現，通過對企業微博用戶群體進行聚類，了解用戶的分佈情況和興趣特徵，為企業提供針對性的顧客關係管理和行銷建議，指導企業更好地發現用戶關注點，提升微博行銷效果。社會治理層面：微博用戶群體的認知態度和行為動機，對於輿情監測和分析是非常重要的（張春華，2012：106）。

根據用戶發佈的文本、傳播行為、用戶偏好、社交關係等資訊，採用支持向量機、Logistic回歸、混合提升演算法（AdaBoost）等方法，對用戶的年齡、性別、職業、情感、偏好和政治傾向，進行聚類、識別和推測，已經取得了很多的研究成果（任薇、邱玉輝，2014；Huffaker, 2010; Poell, Kloet, & Zeng, 2014; Zhao et al., 2012）。

根據社交媒體用戶基於認知層面的自定義標籤和社交網絡，可以實現用戶聚類。Kosinski、Stillwell及Graepel（2013）利用Facebook用戶的likes行為，預測用戶的人格特點、種族、宗教等用戶特徵。熊回香及蔣武軒（2017）依據用戶標籤與關係網絡，採用*K-means*方法，對微博用戶進行聚類。李運華、汪祖柱、葉燕霞及張星星（2016）的論文認為，根據微博用戶的關注數、粉絲數、微博發帖數、微博轉發時間、微博內容、轉發數、評論數、點讚數和微博網址，同樣採用*K-means*方法，可以實現微博用戶的聚類。安璐及周亦文（2020）則從用戶特徵和文本特徵兩個維度，構建指標體系，並採用兩步聚類提取微博用戶特徵。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

利用微博用戶發帖的行為數據、文本資訊甚至表情符，可以推斷用戶的年齡、性別、興趣和情感，實現用戶群體的屬性聚類。但實際應用中，微博用戶聚類經常面臨用戶屬性資料缺失、微博文本資訊不足、用戶行為信息缺失等現實問題：用戶在註冊微博帳號時，會被提示填寫個人資料，但經常會出現用戶資訊填寫不全或不準確的情況。相當部分用戶，僅將微博作為資訊獲取的管道，不定義個人標籤，很少甚至不發佈評論（熊回香、蔣武軒，2017；McGregor & Vargo, 2017）。對於這類用戶，常見的處理方法是直接忽略（李運華、汪祖柱、葉燕霞、張星星，2016；林燕霞、謝湘生，2018）。而該類用戶實際上是互聯網新聞的高頻重度用戶，佔互聯網新聞讀者的比例為15.3%，集中在70後群體，35–44歲佔比達到76.8%（中國互聯網絡信息中心，2017）。因此，只通過用戶微博的評論文本，來進行用戶聚類，會導致大量用戶遺漏，使得聚類結果出現較大偏差。

針對微博用戶資料缺失、文本資訊不足等用戶聚類面臨的實際問題，本文給出了一種基於認知層面的微博用戶聚類方法。微博用戶主觀認知和使用動機的差異，會表現在用戶關注和轉發資訊源的行為上的不同。聚類方法提取微博用戶接收、傳播資訊的認知特性和行為特性，建立關注和轉發兩級聚類機制，構造用戶—關注、用戶—轉發的雙模網絡，得到雙模網絡的鄰接矩陣，對鄰接矩陣進行聯合譜聚類，區分出不同認知屬性的用戶群體，從而實現了微博用戶的聚類。

本研究旨在解決在社交媒體用戶信息缺失情況下的用戶群體劃分的問題。提供的聚類方法，是基於用戶的價值判斷、利益訴求和情感等「隱性態度」方面，不依賴用戶的個人認證信息和發帖的文本信息。這種用戶群體聚類方法，可以提供一個分析角度，來觀察特定的認知群體使用信息源的「隱性認知態度」與「顯性傳播行為」之間的關係。

這種聚類方法，可以壓縮和簡化龐大的微博社交網絡，降低網絡規模。在對用戶聚類的同時，聚類方法對用戶群體關注和轉發的資訊源，也進行了分割。在巨幅的社會網絡連接圖中，該方法為網絡群體的定位和分析，提供了能夠主動聚焦的「放大鏡」，實現在不同分辨率（resolution）下的信息傳播環境的觀測，可以發現進而理解微博龐雜用

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

戶群體中的社群結構以及資訊傳播過程中用戶與資訊源之間的影響關係，研究社交網絡中信息交互模式和集體行為特徵。

文獻綜述

微博用户聚類方法

根據分析的數據是否擁有標記資訊，機器學習分為有標記的「監督學習」和無標記的「無監督學習」。聚類屬於「無監督學習」策略中應用最廣的一種分類方法。當待分析的數據的標記資訊是未知時，聚類可以通過對無標記訓練樣本的學習，完成群體區分，並解釋數據的內在性質及規律。

電子商務行銷領域的推薦機制，是聚類方法最為常見的使用場景。將客戶和商品，基於商品內容的相似性或使用者消費行為的相似性進行聚類，實現個性化精準推薦。

如何對微博用戶進行聚類，同樣是線上社交網絡大數據分析和機器學習的研究熱點。但目前的思路和方法，局限在活躍用戶及顯性行為，非活躍用戶及隱性認知態度是研究的一個盲點。相較於可以直接獲取的微博用戶顯性資料，微博用戶的認知態度和行為動機隱於其後，難以收集、量化和分析。如何分析只看不發言的「潛水用戶」以及傳播行為背後的「隱性態度」，目前的研究是不足的（McPherson, Smith-Lovin, & Cook, 2001）。

Quercia、Capra及Crowcroft（2012）認為Twitter上的議題和情感，會吸引同類用戶聚集在一起。用戶往往關注那些與他們政治意見一致的人們，而很少關注與其政治態度意見相左的資訊。Himelboim、McCreery及Smith（2013）選擇了Twitter上10個有爭議性的政治議題，用以對用戶群體聚類進而分析群體所表達的政治傾向。用戶在對信息源的接收、篩選和再傳遞的過程中，也會將自己的個人認知和情感賦予其中，吸引具有相同觀點和情緒的用戶（Hill et al., 2010）。

從傳播層面來看，用戶作為資訊傳播的中轉環節，其獲取和傳遞資訊，並非被動地完全接受，而是具有相當程度的主動性。用戶通過

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

微博的關注功能，選擇自己認為可信、有用、符合個人價值觀的資訊源，構建用戶自有的獨特的資訊來源管道。關注同一資訊源的用戶群體，由於個體在年齡、性別、職業、學歷、社會經歷和價值觀的差異，對資訊的認同和處理方式也會有所不同。Li、Dittmore、Scott、Lo及Stokowski（2019）比較了Twitter用戶和新浪微博用戶在關注洛杉磯湖人隊動機層面的不同，Twitter用戶側重娛樂和獲取球隊的技戰術指標，而新浪微博用戶更注重獲取新聞資訊和表達粉絲支持。用戶轉發資訊的行為，也能反映出用戶的興趣偏好和價值取向，代表了用戶對資訊和資訊源的認可（許小可、胡海波、張倫、王成軍，2015：111；Hughes et al., 2012; Xu et al., 2011）。

「物以類聚，人以群分」，用戶選擇媒介、參與媒介傳播的行為，實際上已經隱含了用戶的認知和動機。同類用戶社群有如下特點：社群成員之間不一定存在線上和線下的實際聯繫，但關注或轉發相同的資訊源，具有相似的價值認知、資訊獲取的動機和處理的方式。那麼，基於微博用戶關注和轉發資訊源的不同類型，分析用戶從何處接收資訊，認同哪些資訊，就可以有效地劃分用戶的類型。

用戶關注和轉發信息源的行為，實際上已經不自覺將自己劃歸到某一群體中。用戶關注和轉發的信息源越多，呈現出的自身的身份、興趣和認知也就越加清晰。

通過微博用戶關注和轉發資訊的行為，來實現微博用戶的聚類，可以有效解決用戶屬性數據、文本數據和行為數據不足的現實問題，可以合理推斷用戶的「隱性態度」。這是本文的研究思路。

微博用户關注和轉發行為

微博用戶的多元分層結構，使得其資訊需求和傳播行為呈現出多樣化。目前，對於微博用戶關注和轉發行為的相關研究，大多從話題推薦、微博行銷、輿情監控等實際應用層面來展開分析。

Phelan、McCarthy、Bennett及Smyth（2011）通過分析Twitter用戶訂閱的RSS新聞資訊源，提取出微博用戶感興趣的微博主題，進而實

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

現新聞話題的智能推薦。智能精準推薦可以採用改進的協同過濾方法，向微博用戶推薦值得關注的好友（姚彬修、倪建成、于蘋蘋、李淋淋、曹博，2017）。林燕霞與謝湘生（2018）使用社會認同理論，解釋微博用戶的關注行為，實現用戶所屬群體的分類，提取群體的特徵屬性。Yun、Pamuksuz 及 Duff（2019）認為 Twitter 用戶個性與關注的品牌的特徵之間存在一致性，能夠應用於品牌推廣領域。

對微博用戶轉發行為的研究主要集中在以下方面：分析影響用戶行為的因素和預測用戶的轉發行為（唐曉波、羅穎利，2017；Boyd, Golder, & Lotan, 2010; Peng et al., 2011; Petrovic, Osborne, & Lavrenko, 2011; Suh et al., 2010）。陳姝、竇永香及張青傑（2017）考慮了微博語義特徵、數據形式、用戶特徵和用戶交互等因素，探究了這些因素如何作用於微博用戶主觀層面，進而影響其傳播行為。丁緒武、吳忠及夏志傑（2014）考慮情感因素對用戶轉發行為的影響。Abdullah、Nishioka、Tanaka 及 Murayama（2017）分析了用戶在社交媒體上轉發災害資訊，是兼具了利他和利己的雙重考慮。

上述研究說明，微博用戶受到微博文本、社交關係和自身主客觀條件等諸多因素的影響，兼具理性判斷和情感驅使，在關注和轉發資訊源行為會出現群體間的差異。

用戶關注某個資訊源，可以認為是用戶對該資訊源的一種訂閱行為。新浪微博的轉發設計，使得用戶在轉發微博信息的同時，可以選擇是否對微博信息發表評論。轉發和評論行為具有天然的一致性。

從用戶動機角度來看，關注行為關聯到用戶個人層面的信息獲取，轉發行為關聯到用戶社交層面的自我表達。從認知角度來看，關注行為是用戶對信息處理的輸入端，轉發行為是用戶對信息處理的輸出結果。

因此，從用戶的動機和認知角度，需要考慮用戶對關注和轉發資訊源行動之間的差異。這些差異，從不同的側面更為完整地展示了微博用戶的整體畫像。這是微博用戶聚類方法中綜合考慮關注和轉發用戶行為的依據。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

社會網絡分析方法

對於用戶的聚類分析，可以採用社會網絡的分析方法(Barsade, 2002; Castellano, Fortunato, & Loreto, 2009; Gomez et al., 2013)。傳統的聚類方法，由於用戶屬性的缺失，難以構造出網絡的拓撲結構。利用微博用戶關注和轉發資訊的行為，可以建立關係的連接，從而形成網絡：網絡中的點代表受眾和資訊源；網絡中點與點之間的連接線，代表用戶與資訊源間的關注和轉發關係。通過全面描繪該網絡，能夠發現網絡中的特定模式，追蹤資訊在網絡中的流動方式，發現特定的連接關係和對應的子網絡對用戶的影響。

苑衛國、劉雲、程軍軍及熊菲(2013)以新浪微博中普通用戶和名人用戶兩類用戶的雙向關注網絡為研究對象，探討了微博資訊傳播網絡的結構、路徑及其影響因素。袁園、孫霄凌及朱慶華(2012)利用網絡分析方法，進行整體微博信息網絡分析、內部子結構分析和角色位置分析，進而從微博關注數據中挖掘用戶關注對象的分佈及對象間的關聯性。

微博用戶的資訊關注和轉發網絡，是社會網絡的一種特定的形式。Sun等(2009)使用RankClus聚類框架，實現異質網絡的用戶聚類。譜聚類方法，是聚類和社區群體發現的經典算法，效率較高。社會網絡分析是網絡結構的分析，它的分析單位是連接關係。這種連接關係，可以轉化為網絡的鄰接矩陣。譜聚類方法的「譜」，指的是鄰接矩陣的特徵值和特徵向量。譜聚類方法，是通過鄰接矩陣的特徵值和特徵向量，將代表用戶的網絡中的點，劃歸到某一特定類別。

建立微博用戶資訊網絡，並對其進行二部圖的譜聚類，這是本文研究微博用戶聚類的具體方法。

微博用戶的聯合譜聚類方法

聯合譜聚類方法原理

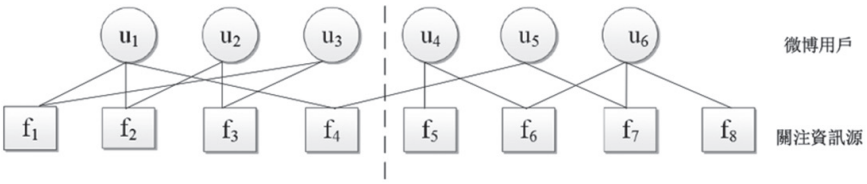
社會學利用網絡科學理論來描述社會網絡。社會網絡中的一切實體或事件均可稱為行動者(actor)，在網絡圖中以節點呈現。模(mode)

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

定義為行動者的集合，模數表示行動者集合類型的數目。由一類行動者集合與另一類行動者集合之間的關係構成的網絡，稱為雙模網絡（two-mode network）。利用微博用戶關注和轉發資訊的行為，可以建立雙模網絡。從圖論的角度來看，雙模網絡是典型的二部圖結構。

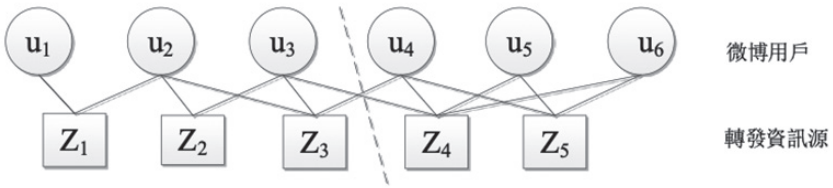
微博用戶關注他人的微博號，轉發他人的微博資訊，建立資訊接收和分發的聯繫，形成了兩個雙模網絡：用戶—關注網絡和用戶—轉發網絡。每一個模代表一類實體，分別是：微博用戶和關注資訊源、微博用戶和轉發資訊源。

圖一　微博用戶與關注資訊源的雙模網絡



如圖一所示，該網絡是一個6個微博用戶關注8個資訊源的雙模網絡。用戶與轉發資訊的雙模網絡，如圖二所示。與關注資訊源不同，微博用戶有時會轉發自己的資訊，不是嚴格意義的二部圖。簡化起見，我們將轉發自身資訊的用戶，轉換為另一個鏡像用戶，依然保留二部圖結構。

圖二　微博用戶與轉發資訊源的雙模網絡



在雙模網絡中，一個模集合的聚類形成，是由於這個模集合中的個體共同與另一個模集合中的集合相連接。在微博用戶—關注資訊源、用戶—轉發資訊源的雙模網絡中，興趣相同的用戶更傾向於接收相似的資訊，認知相同的用戶更傾向於轉發相似主題和內容的資訊。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

對於微博用戶的關注和轉發行為的研究，目前的處理方式，或者是將關注和轉發不加區別地統一劃歸到微博用戶行為類別（田占偉、王亮、劉臣，2015），或者是割裂開來，將轉發劃歸到微博用戶行為類別，而排除掉關注行為（劉繼、李磊，2013）。只考慮轉發行為的處理方式，是出於實用層面的考慮，更關心資訊大規模傳播的後果，目的是誘導或者限制用戶對相關資訊的傳播：促使社交媒體話題的傳播擴散，可以帶來網絡營銷的商業價值；採取有效的輿論管控和輿論引導，可以限制網絡群體性事件的廣泛傳播。

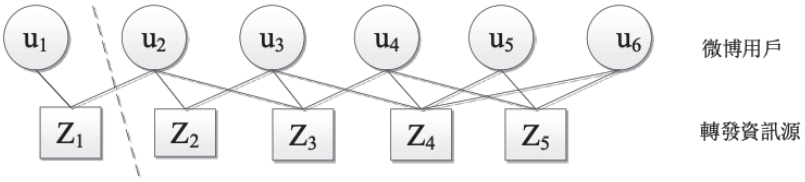
綜合考慮關注和轉發兩種不同形式，是基於微博用戶認知的考量：某些用戶雖然主動接觸大量資訊源，但並不熱衷表達自己的觀點，很少評論和轉發資訊；有些用戶雖然關注同一資訊源，但認知態度可能截然相反；還有的用戶關注大量不同類型資訊源，發帖均為轉發資訊。這類用戶往往出於商業行為，成為各類資訊的中轉站，對資訊沒有任何個人的意見和態度。

根據網絡的整體拓撲結構，對雙模網絡中二部圖的頂點進行分割，劃分出不相交的集合。圖一和圖二中的虛線，表示了二部圖的分割。這種劃分，可以同時得到用戶和關注資訊源的集合、用戶和轉發資訊源的集合：通過微博用戶與關注、轉發資訊源的聯結關係，可以將微博用戶$\{u_1, u_2, u_3, u_4, u_5, u_6\}$分為$\{u_1, u_2, u_3\}$和$\{u_4, u_5, u_6\}$兩類，將關注資訊源$\{f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8\}$分為$\{f_1, f_2, f_3, f_4\}$和$\{f_5, f_6, f_7, f_8\}$兩類，將轉發資訊源$\{z_1, z_2, z_3, z_4, z_5\}$分為$\{z_1, z_2, z_3\}$和$\{z_4, z_5\}$兩類，其中$\{u_1, u_2, u_3\}$類用戶，關注的資訊源類是$\{f_1, f_2, f_3, f_4\}$，轉發的資訊源類是$\{z_1, z_2, z_3\}$；$\{u_4, u_5, u_6\}$類用戶關注資訊源類是$\{f_5, f_6, f_7, f_8\}$，轉發資訊源類是$\{z_4, z_5\}$。

二部圖採用經典的最小割集的方法，可以得到高效求解，但所得的解往往是不平衡的，有的集合可能只包含很少的頂點，甚至只有一個頂點。如果將微博用戶與轉發資訊源都進行最小化切割，因為微博用戶u_1只轉發一個資訊源z_1，微博用戶的聚類結果為：$\{u_1\}$和$\{u_2, u_3, u_4, u_5, u_6\}$，轉發資訊源聚類結果為：$\{z_1\}$和$\{u_2, u_3, u_4, u_5\}$，如圖三所示。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

圖三　微博用戶與轉發資訊源的最小化切割



與圖二的分割比較，圖三的最小化切割沒有考慮到用戶與資訊源的特定的轉發關係，對於分析用戶群體的認知和傳播行為沒有實際應用價值。源於圖劃分問題的譜聚類方法就是尋找一種切割方法，可以將圖中的頂點分割成若干不相交的集合，在兼顧集合規模的前提下，儘量切割最少的邊（Von Luxburg, 2007）。聯合譜聚類，採用了規範化割集準則，在保證切割邊最小化的同時，避免了如圖三所示的只包含一個成員的小聚類群體的現象。採用聯合譜聚類方法，可以有效解決切割不平衡的問題，並能同時得到用戶及其關注資訊源、用戶及其轉發資訊源的聚類。

K-means 聚類方法，是聚類中最常用的算法。其基本原理如下：設定聚類的個數為 K，隨機確定 K 個初始點為質心，計算各頂點到這 K 初始點之間的距離。通過距離計算，為每個頂點找到最近的質心，將其分配到該質心對應的類別中，同時該類質心的位置會根據新分配進來的頂點位置進行調整。這個過程反覆迭代，直到誤差收斂或達到最大運算次數。*K-means* 聚類方法效率快，但聚類的對象只針對同類頂點，只能實現單一的用戶群聚類或者資訊群聚類。

譜聚類方法原理如下：由雙模網絡生成對應的鄰接矩陣，採用多維向量方法，對鄰接矩陣進行降維。將鄰接矩陣的特徵向量分量作為新的參考系，把網絡中的所有頂點映射到一個新空間，使得原始網絡在新的空間中更容易被劃分（Charu, 2016, p. 93; Tang & Liu, 2010; Trivedi & Dey, 2016）。

根據用戶名稱和用戶關注集，建立以用戶名稱為行，用戶關注集元素為列的用戶—關注鄰接矩陣 A_g，建立以用戶名稱為行，用戶轉發

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

集元素為列的用戶—轉發鄰接矩陣A_z。以A_g每行元素的和，作為對角矩陣的對角元素值，得到用戶關注連接度矩陣D_{ug}。計算A_g每列的元素的和，得到對角矩陣D_{vg}，即關注集連接度矩陣。依此方法，得到用戶轉發連接度矩陣D_{uz}、轉發集連接度矩陣D_{vz}。

規範鄰接矩陣A_g、A_z，並對規範後的鄰接矩陣$\tilde{A}_g$，$\tilde{A}_z$進行奇異值分解。

鄰接矩陣的規範方法為：

$$\tilde{A}_g = D_{ug}^{-1/2} A_g D_{vg}^{-1/2} \tag{1}$$

$$\tilde{A}_z = D_{uz}^{-1/2} A_z D_{vz}^{-1/2} \tag{2}$$

對規範後的鄰接矩陣進行奇異值分解：

$$\tilde{A}_g = U_g \Sigma_g V_g^{\mathrm{T}} \tag{3}$$

$$\tilde{A}_z = U_z \Sigma_z V_z^{\mathrm{T}} \tag{4}$$

將鄰接矩陣的特徵向量分量作為新的參考坐標系，把網絡中的節點映射到新空間中，得到特徵向量$S_g^{(u)}$，$S_g^{(v)}$，$S_z^{(u)}$，$S_z^{(v)}$。

$$S_g^{(u)} = U_{g(1:k)}，S_g^{(v)} = U_{g(1:k)} \tag{5}$$

$$S_z^{(u)} = U_{z(1:k)}，S_z^{(v)} = U_{z(1:k)} \tag{6}$$

其中k值是待劃分的用戶類別數，可採用肘點法（elbow method）獲取聚類數目。

由式（5）、（6），得到新空間的聯合聚類指標集矩陣Z_g，Z_z

$$Z_g = \begin{bmatrix} D_{ug}^{-1/2} S_g^{(u)} \\ D_{vg}^{-1/2} S_g^{(v)} \end{bmatrix} \tag{7}$$

$$Z_z = \begin{bmatrix} D_{uz}^{-1/2} S_z^{(u)} \\ D_{vz}^{-1/2} S_z^{(v)} \end{bmatrix} \tag{8}$$

對聯合聚類指標集矩陣Z_g，Z_z完成二級處理和*K-means*聚類。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

微博用户譜聚類的算法

微博用戶譜聚類的算法流程如下：

(1) 爬取微博用戶資訊，獲取用戶基本屬性、傳播行為和資訊傳播路徑。內容包括：用戶頁面的用戶認證資訊、關注資訊源、轉發資訊源和轉發內容；用戶關注和轉發頁面的關注者或轉發者名稱、粉絲數和個人認證資訊。其中的資訊源信息直接用於用戶聚類，而用戶和資訊源認證信息、轉發內容用於後續聚類效果的檢驗。

(2) 建立用戶關注集和用戶轉發集。收集用戶關注資訊源和轉發資訊源數據，合併有重複關注和轉發的對象，得到用戶關注集和轉發集。對於受眾的關注集和轉發集數據，考慮了三種不同的處理思路：並集處理、交集處理和二級處理。並集處理是將關注資訊源和轉發資訊源合併，剔除掉重複的資訊源，再進行聚類分析；交集處理是在用戶關注的資訊源中選擇其中的轉發資訊源，即選擇關注集和轉發集的交集。二級處理，則是先進行關注資訊源聚類，在此聚類基礎上，按照轉發資訊源進行二次聚類。

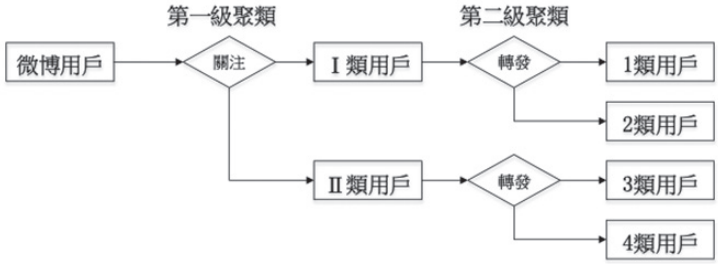
需要注意的是，二級處理要考慮到總聚類數和各級聚類數之間的關係。如式 (9) 所示，總聚類數目 k 與二級處理中的聚類數目 k_1、k_2 為乘積關係。

$$k = k_1 \times k_2 \tag{9}$$

這三種處理方式是源於用戶對資訊源的使用動機來進行區分。分析大量用戶關注和轉發行為，可以推斷：關注資訊源來自用戶對於職業、生活和娛樂等方面資訊獲取的需要。轉發往往表示用戶在認知上對轉發資訊的認可或否定。而關注該資訊源且轉發該資訊源，可以更為明確地反映用戶的認知情況。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

圖四　微博用戶二級聚類機制



如圖四所示，採用資訊源的二級聚類方法，分別考慮了用戶對資訊源在關注和轉發上不同的使用目的和認知傾向，對於數據的運用更加充分，可以有效識別用戶群體，更為準確定位影響用戶群體的資訊源。

三種處理方式，對應了不同的目標用戶群體。在對使用者行為資料進行統計分析的基礎上，作如下推測：適合交集處理的用戶群體，往往出於娛樂休閒動機，會關注大量的社交名人和娛樂明星，轉發明星動態，關注的資訊源數量在500以上。普通微博用戶的關注上限是2,000，這類用戶群體經常會開通微博會員，提高關注人數上限；適合並集處理的用戶群體，偏重社會交往目的，不將微博作為主要的信息獲取渠道，更關注生活層面，會在微博上記錄自己的日常經歷和感受，關注線下真實的親朋好友，轉發朋友關注的資訊信息。這類用戶群體關注或者轉發的信息源有限，關注數不超過100個，需要並集處理，才能較為準確獲取用戶的認知；適合二級處理的用戶群體，有明確的信息獲取動機。這類用戶群體對關注和轉發的信息源會有意識的挑選，會關注行業專家和意見領袖，關注數一般不超過200個，中位數在60個左右。

實際應用中，還需要考慮數據規模和處理效率的影響。當用戶關注資訊源和轉發資訊源數量較少時，可以考慮採取並集處理的方式；當用戶關注資訊源和轉發資訊源數量太大、設備處理性能受限的情況

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

下，可以考慮採取交集處理的方式；在資訊源數量足夠，並且處理性能滿足需求時，優先考慮二級處理的聚類方式。

在上述三種聚類處理方式的基礎上，還可以將並集作為新的關注集，交集作為新的轉發集，得到新的混合二級聚類機制。在後續的實例分析中，比較了這幾種處理方式的性能。

(3) 建立用戶—關注鄰接矩陣、用戶—轉發鄰接矩陣、用戶關注連接度矩陣、關注集連接度矩陣、用戶轉發連接度矩陣、轉發集連接度矩陣。

對建立的用戶關注集和用戶轉發集，分別應用聯合譜聚類算法，得到聯合聚類指標集矩陣Z_g、Z_z。聯合聚類指標集矩陣Z_g、Z_z、的第一列，無法提供有效的聚類資訊。從第二列開始，通過*K-means*聚類算法，完成用戶與資訊源的聯合聚類。

與直接使用*K-means*聚類方法比較，譜聚類算法的優勢在於能夠同時得到用戶群和及其關聯的資訊群，構建資訊源和用戶的異構的信息傳播網絡，可以觀察到什麼樣的用戶關注哪類資訊源，為後續分析用戶認知傾向和資訊源對用戶認知的影響，設定明確的群體傳播環境。

微博用户聚類效果的性能度量

對於聚類方法，需要有性能指標來評估聚類結果的準確性。聚類任務屬於「無監督學習」，訓練樣本的標記資訊是未知的。這種情況，如何評估聚類效果的好壞？

聚類性能度量可以使用「外部指標」和「內部指標」兩種方式。將聚類結果與某個參考模型或參考樣本進行比較，稱之為「外部指標」度量；直接利用聚類結果，不依賴任何外部標註資訊或模型，稱之為「內部指標」度量。

「內部指標」的評價原理是認為好的聚類情況下同一類的樣本彼此近似，不同類的樣本會盡可能地不同。因此，比較聚類結果中同類用戶的「類內相似度」和不同類用戶的「類間相似度」，可以衡量聚類的效果。聚類效果越好，則「類內相似度」越高，「類間相似度」越低。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

內部指標可以採用DB指數（Davies-Boudin Index）來進行度量。

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{cen}(\mu_i,\mu_j)}\right) \quad (10)$$

其中：聚類結果劃分為$C=\{C_1, C_2\cdots, C_k\}$；$k$表示聚類數目；$avg(C_i)$定義為聚類第$i$類內部平均距離；$avg(C_j)$定義為聚類第$j$類內部平均距離；$d_{cen}(\mu_i, \mu_j)$表示第$i$類中心點$\mu_i$和第$j$類中心點$\mu_j$的距離。

DB指數值越小，表示「類內相似度」高，「類間相似度」低，聚類效果越好。

表一　雙模網絡轉化的鄰接矩陣

	z_1	z_2	z_3	z_4	z_5
u_1	1	0	0	0	0
u_2	1	1	1	0	0
u_3	0	1	1	1	0
u_4	0	0	1	1	1
u_5	0	0	0	1	1
u_6	0	0	0	1	1

以圖二的微博用戶與轉發資訊源的雙模網絡，介紹DB指數計算方法。根據雙模網絡，得到如表一所示的鄰接矩陣。通過鄰接矩陣，得到u_1的向量為[1,0,0,0,0]，u_2的向量為[1,1,1,0,0]，計算二者向量的歐氏距離，得到樣本u_1和u_2差異度量值。由此，可以計算出樣本間的平均距離、聚類中心點和類間距。

假設圖二有兩種聚類方式：聚類方式1，分為{u_1，u_2，u_3}和{u_4，u_5，u_6}兩類；聚類方式2，分為{u_1，u_2，u_4}和{u_3，u_5，u_6}兩類。由式（10）得到聚類方式1的DB指數值為1.456，聚類方式2的DB指數值為4.129。這説明第一種聚類方式是優於第二種聚類方式。

在實際應用中評估聚類效果，可以綜合應用兩種評估方式：聚類算法設計和參數調節階段，可以選擇小樣本微博用戶，對其進行人工標註類別，得到「外部指標」，評估聚類效果，便於算法的改進和參數的調整。當算法投入大規模用戶聚類的應用環節，無法進行人工判定時，就需要應用「內部指標」，作為聚類度量標準。同時，「內部指標」還可以作為聚類過程中的優化目標，指導聚類符合要求。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

微博用戶聚類方法的具體應用實例

本節以2018年新浪微博用戶關於「湯蘭蘭事件」的討論，作為具體對象，來驗證聯合譜聚類方法的可行性和有效性，並結合微博用戶的實際情況，對方法作進一步的改進。

湯蘭蘭事件概況和數據收集說明

2008年10月，14歲的少女湯蘭蘭（化名）向當地警方舉報，稱自己從6歲開始被父親、爺爺、親屬和鄉鄰等十餘人強姦、輪姦，時間長達7年。2010年10月，黑河市中級法院判決了湯案，判處包括湯蘭蘭父母在內的11人無期徒刑或5–15年有期徒刑不等。2017年末，湯蘭蘭母親出獄，要求重審此案。

「湯蘭蘭事件」在網絡傳播和熱議，可劃分為三個階段。第一階段：2018年1月31日，《澎湃新聞》和《新京報》的兩篇報導，列舉了案件中的諸多疑點，將湯蘭蘭的戶籍資訊部分公開，提出要「尋找湯蘭蘭」，引起了廣泛的爭議。「我不是謙哥兒」、「午後的水妖」等微博大V，質問媒體動機，其觀點得到微博用戶廣泛的認同並迅速擴散，成為微博輿論場的引爆點，「陰謀論」、「媒體審判」等言論成為了微博用戶的主要觀點，澎湃記者和新京報評論員受到網上的各種批評和攻擊。第二階段：2018年4月26日，中央電視台法治節目播出的「湯蘭蘭案再關注」，針對涉案證據和疑點進行了報導，重新喚起了公眾的關注。第三階段：2018年7月27日，黑龍江省高級人民法院宣佈了湯蘭蘭案再審的審查結果，駁回「湯蘭蘭案」原審被告人的申訴。在持續近半年的「湯蘭蘭事件」中，幾百萬微博用戶，圍繞探求疑案真相、受害人隱私保護、媒體職業操守等議題，觀點針鋒相對，展開了激烈的討論。

將「湯蘭蘭事件」作為微博用戶聚類的分析對象，有如下幾個理由：選擇「湯蘭蘭事件」，是基於該事件傳播的廣度和深度的考慮。「湯蘭蘭事件」是2018年中國大陸社交媒體傳播的熱點事件，傳播範圍廣，相關話題的微博經常有百萬級的閱讀量、萬條的評論帖。話題獲得了大量轉發，傳播層級多，有的單條微博的轉發傳播的最大深度高達15

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

層(知微數據，2018)。作為聚類的對象，該實例有數量豐富、用戶類型多樣的數據樣本可供分析。選擇「湯蘭蘭事件」，同時也是基於觀點的平衡性和數據準確性的考慮。「湯蘭蘭事件」中，官方媒體、商業媒體、自媒體和普通微博用戶能夠發表不同的觀點，並展開討論。雖然有的觀點受到了微博用戶的批評甚至攻擊，但基本上不同觀點和意見能夠平等、開放地出現在微博上。近年來，這種公共話題的多種觀點呈現的現象已經不多見了。如果獲取的數據都是單一的評論立場和壓倒性的觀點，則已經自然實現了用戶的篩選和聚類，無法有效衡量聚類算法的性能。對於「湯蘭蘭事件」，微博用戶群體沒有出現一面倒的主導性觀點，有利於檢驗聚類方法的效度；微博上對「湯蘭蘭事件」的討論，不同於群體性事件和商業熱搜事件，較少受到輿情管控和商業營銷因素的干擾，提取的評論資訊和用戶資訊較為準確和完整，有利於檢驗聚類方法的信度。

Conover等(2013)分析了全美範圍內Twitter用戶中關於佔領華爾街事件的關鍵字，發現跨州傳播的帖子中被強調的資訊主要涉及了該運動的核心框架以及新聞報導主題。Twitter在集體敘事框架的構建過程中，起到關鍵作用。微博公共事件討論中的代表性觀點，也可以通過議題關聯與關係建構方式，進行媒體網絡議程分析(黃敏，2020)。

表二　代表性觀點的認知框架

觀點	認知框架		
	問題定義 A	因果解釋 B	策略或呼籲 C
1	事件存疑A1	冤案B1	尋找湯蘭蘭，追求真相C1
2	事件存疑A1	冤案B1	輿論監督、尋找疑點證據C2
3	侵犯隱私A2	職業操守B2、 專業能力不足B3	追責相關記者和媒體C2
4	事件存疑A1， 但侵犯隱私A2	專業能力不足B3、 媒體監督困境B6	回歸真相，聚焦案件本身C1
5	侵犯隱私A2， 鐵案A3	媒體和律師企圖翻案B4	輿論干擾司法，媒體追責C2
6	法治進程中的 客觀現象A4	法律框架內理性分析和 良性互動B5	良性互動、共同尋求真相C3

「湯蘭蘭事件」的代表性觀點，按框架效果可參見表二所示。各種議題的代表性文章，參見表三。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

表三　各種議題的代表性文章

議題類型	文章標題
1	〈尋找湯蘭蘭〉
	〈女孩被全家「性侵」被告喊冤10年還原案件始末〉
2	〈湯蘭蘭案的無盡疑點：口供、黃碟、陰陽B超單〉
	〈湯蘭蘭案最詳細調查在此〉
3	〈澎湃新聞記者王樂，受害者的人血饅頭好吃嗎〉
	〈湯蘭蘭不該找，媒體人該檢討〉
4	〈放過湯蘭蘭也放過記者，請回歸案件本身〉
	〈怎麼評價澎湃對湯蘭蘭一案的報導？為什麼審判媒體是錯的？〉
5	〈記者和媒體可以監督公權，那誰來監督記者和媒體？〉
	〈官方回應「湯蘭蘭」案：其母串聯他人借媒體炒作 企圖翻案〉
6	〈「湯蘭蘭案」終須回歸法治管道〉
	〈面對「尋找湯蘭蘭」，我們應該怎麼辦〉

本研究共收集了新浪微博用戶在2018年1月19日到2018年7月27日期間發佈的評論資訊，包括有代表性的145篇報導，相關發帖575,298條。刪去無意義的文本和表情符，保留10個字數以上的有效評論，得到文本評論共計261,167條，其中第一階段評論138,852條，第二階段18,103條，第三階段104,212條。

聯合譜聚類的應用

通過文本分析，選擇42名有代表性的用戶。我們先以這42名評論用戶作為聚類對象，應用聯合譜聚類方法，完成用戶和關注資訊源、用戶和轉發資訊源的聚類，並驗證微博用戶聯合譜聚類方法的可行性，評估三種聚類方式的效果。進而，將聚類方法推廣應用到多階段參與的6,737個用戶中，驗證演算法的有效性和魯棒性（robustness）。

選擇42名用戶的標準：用戶發佈的觀點涵蓋了表二的六種主要觀點；用戶對湯蘭蘭事件有較高的關注度，在事件發生的三個階段都全程參與評論或轉發。

抓取的微博用戶的關注和轉發的資訊源，均有相當部分是重合的。這說明，對微博用戶進行聯合譜聚類方法，在理論上是可行的。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

隨著用戶群體數量的增加，社交網絡也會隨之擴張，帶來用戶關注和轉發的資訊源數量急劇增加，鄰接矩陣A_g、A_z會成為大規模的稀疏矩陣。因此，基於傳播層面和算法效率的考慮，只選擇粉絲數在10萬以上的資訊源進行分析。粉絲數10萬以上的資訊源，總體數量較為穩定。2013年的統計資料顯示：新浪微博和騰訊微博中，粉絲數10萬以上的大V用戶超過1.9萬個，100萬以上的超過3,300個（張翼、李培林、陳光金，2013：215）。通過刪除重複發言用戶和選擇大V資訊源的方法，可以提升微博用戶聚類方法的效率，減少運算規模，降低運算成本。

除去開通微博自動關注的微博小秘書、微博客服等微博號，並刪除重複資訊源，取粉絲關注數為10萬以上的大V資訊源4,351個，去重後得到的資訊源3,143個。同樣處理，轉發資訊源取自用戶頁面的前二十頁，得到轉發獨立資訊源2,888個。建立相應的鄰接矩陣：微博用戶—關注資訊源鄰接矩陣為42行3,143列的矩陣，42為用戶數，3,143為用戶關注的資訊源數；微博用戶—轉發鄰接矩陣為42行2,888列的矩陣，42為用戶數，2,888為用戶轉發的資訊源數。

資訊源的處理方式可以分為：並集處理方式、交集處理方式和二級處理方式。並集處理方式：將關注資訊源和轉發資訊源合併考慮，作為用戶接觸的資訊源。由於關注資訊源和轉發資訊源二者存在重疊部分，因此得到資訊源並集有4,972個。交集處理方式：分析關注資訊源中有多少是用戶轉發的資訊源，將這部分資訊源挑選出來，得到資訊源交集有1,059個。二級處理方式：先使用關注資訊源3,143個資訊源進行一級聚類，然後在聚類的基礎上，使用轉發資訊源2,888個，進行二次分類。

由於微博的設置，微博用戶可以顯現的關注資訊源只有前50頁，雖然微博過濾掉了部分廣告資訊，但目前獲得的關注資訊源仍然是不完整的，會帶來後續聚類分析的偏差。通過將關注資訊源和轉發資訊源的合併處理，作為用戶的資訊源，可以有效地提高分類的準確性。而交集處理方式，能夠更加凸顯出用戶的認知傾向。二級處理則是綜合考慮了用戶對資訊源在關注和轉發上的不同使用目的。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

三種聚類方式的區別，體現在資訊源的處理方式上，在譜聚類算法層面，則基本上是一致的。這裡，以處理較為繁瑣的二級聚類方式為例，給出聚類方法的操作過程。

根據式（4），規範鄰接矩陣A_g、A_z，並對規範後的鄰接矩陣$\tilde{A}_g$，$\tilde{A}_z$進行奇異值分解，由式（5）和式（6），得到特徵向量：

$$S_g^{(u)} = U_{g(1:2)}，S_g^{(v)} = V_{g(1:2)}，S_z^{(u)} = U_{z(1:2)}，S_z^{(v)} = V_{z(1:2)}$$

根據式（7）和式（8），得到聯合聚類指標集矩陣Z_g，Z_z

$$Z_g = \begin{bmatrix} D_{ug}^{-1/2} S_g^{(u)} \\ D_{vg}^{-1/2} S_g^{(v)} \end{bmatrix}，Z_z = \begin{bmatrix} D_{uz}^{-1/2} S_z^{(u)} \\ D_{vz}^{-1/2} S_z^{(v)} \end{bmatrix}$$

對Z_g和Z_z進行*K-means*聚類，劃分用戶類別。

聚類數目的選取

如何確定最佳的聚類數目，使得k值盡可能接近真實的類別數目，是聚類分析的重要問題。常見的方法是通過輪廓系數或肘點，作為確定聚類數目的指標。這裏選擇肘點法，出於以下因素考慮：計算量小、理解直觀。

首先設定聚類誤差和*SSE*，評估聚類效果的好壞。

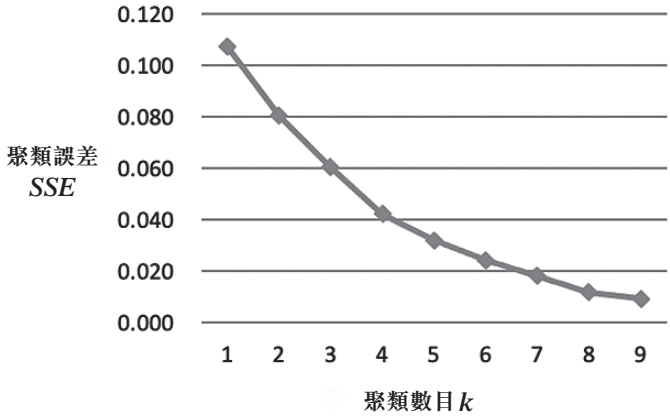
$$SSE = \Sigma_{i=1}^{k} \Sigma_{p \in Ci} |p - m_i|^2 \qquad (11)$$

其中C_i是第i類，p是C_i中的樣本點，m_i是C_i中所有樣本的均值。

肘點法的基本思想：隨著聚類數k的增大，樣本劃分更加精細，每類的聚合程度會逐漸提高，聚類誤差和*SSE*會隨之變小。當k遠小於實際聚類數時，隨著k的增大，會大幅增加每個類別的聚合程度，故*SSE*的下降幅度會很大；當k接近真實聚類數時，所得到的「邊際效用」收益*SSE*會趨於平緩。k和*SSE*對應關係的折線圖，類似手肘的形狀，而肘點對應的k值就接近實際聚類數。這裡將*SSE*的下降幅度高於10%的臨界點位置定義為肘點，由此確定聚類的數目k。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

圖五　聚類數目與聚類誤差的對應關係圖



如圖五所示，k為5的時候，聚類誤差下降幅度低於10%，因此合理的聚類數目應為4。

實際應用中，在肘點法確定最佳聚類數目基礎上，還要考慮是否採取二級聚類的情況。若採取二級聚類，如式(9)所示，需要將聚類數目k因式分解為二級處理中各級聚類數目k_1、k_2，可將聚類值k就近調整到便於整數分解的數值(例如4、6、8、9……等)。

聚類結果分析

確定了最佳聚類數目4，採用二級處理原則，則將每級聚類數目分解為2。按照譜聚類方法，對42名微博用戶進行聚類分析，k值取為2，奇異值分解後的向量U_g，V_g，U_z，V_z均取前兩列。

由式(7)計算，得到關注聚類指標集矩陣Z_g

$$D_{ug}^{\frac{1}{2}} S_g^{(u)} = \begin{bmatrix} -0.01532129 & 0.01532129 \\ -0.01532129 & 0.00448625 \\ -0.01532129 & -0.01347884 \\ \vdots & \vdots \\ -0.01532129 & -0.00398617 \\ -0.01532129 & 0.00189992 \\ -0.01532129 & -0.00680073 \\ -0.01532129 & 0.00166645 \end{bmatrix}$$

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

矩陣Z_g的第一列值相同，表示將所有的用戶和關注資訊源劃分為同一大類，沒有為用戶聚類提供分割的有效指標值。因此，從矩陣的第二列數值進行聚類，由*K-means*聚類方法，可以將用戶分為兩類。兩類用戶的關注資訊源，也同時得到了聚類。

觀察42名微博用戶的原始資訊，進行內容分析，可以發現：第一類用戶關注資訊源，類型多為生活、明星、影視、娛樂類和自媒體，而第二類用戶關注資訊源，內容集中在法律、時政和經濟類資訊。建立的用戶—轉發矩陣，聯合譜聚類後，同樣將用戶分為兩類。組合關注聚類和轉發聚類過程，建立用戶兩級聚類模型。二級聚類，將42名用戶劃分為四類。

根據微博用戶已有的認證資訊和評論資訊，得到如表四所示的42名微博用戶的用戶畫像。

表四　二級聚類的用戶畫像

類型	用戶集	用戶關注類型
1類	8, 9, 18	法律行業人士
2類	6, 27, 28	簽約自媒體、頭條文章作者
3類	1, 2, 7, 10, 15, 19, 21, 23, 25, 32, 34, 37, 38, 39, 41, 42	明星、娛樂、體育、遊戲
4類	3, 4, 5, 11, 12, 13, 14, 16, 17, 20, 22, 24, 26, 29, 30, 31, 33, 35, 36, 40	時政、財經、生活

分析這42名用戶的微博認證資訊，發現聚類劃分的群體能夠準確體現出用戶間的類別差異，聚類效果是可信的。第一類用戶，關注法律類資訊，身份是涉及到法律各部門的政法幹部、律師和網警。第二類用戶，關注時政新聞，關注和轉發資訊數量較多，類型也較為多元，身份為簽約自媒體和頭條文章作者。第三類用戶，是關注娛樂和時事類資訊的年輕人，轉發的大都是娛樂新聞，佔總轉發訊息源的87%。第四類用戶，喜歡關注和轉發財經類和時政類新聞，是教育程度較高、30歲以上的職場專業人士。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

由於樣本較少，42名微博用戶身份的確定，可以採取人工標註的方式。將用戶聚類的結果，與人工判定的分類標註，進行比對，計算出準確率和查全率，檢驗算法的有效性。

人工標註的具體方法如下：兩名標註人員，通過閱讀42名微博用戶的個人認證信息、發帖文本內容、關注和轉發信息源、社交關係等信息，獨立完成微博用戶身份的確定。然後，按照劃分四類的要求，進行用戶歸類。若出現標註和歸類中的爭議情況，則由第三名監督人員進行協商和判定。42名微博用戶中的大部分，有較為詳盡的個人認證信息，較難出現不易判定的情況，兩名標註人員的一致性達到95.2%。出現爭議的有二例，均是由於用戶身份跨界，出現了劃歸到第一類還是第二類的爭議。最後，判定為依據用戶的職業身份來進行區分。

使用「外部指標」來檢驗聚類的準確性，表五給出了僅使用關注關係、僅使用轉發關係、關注和轉發關係並集處理、關注和轉發關係交集處理、關注和轉發關係二級處理、關注和轉發關係混合二級處理六種方式的聚類效果。為了便於比較，僅使用關注關係和僅使用轉發關係的聚類數目也取為4。

表五　用戶譜聚類方法性能比較表

	僅使用關注關係	僅使用轉發關係	並集處理	交集處理	二級處理	混合二級處理
查準率	42.86%	28.57%	69.04%	100%	88.09%	100%
查全率	64.28%	52.38 %	83.33%	38.09%	80.95%	97.62%

比較六種處理方式的性能指標，可以發現，使用混合二級的聚類方式，性能上表現最好。僅使用關注關係和轉發關係，聚類性能較差。觀察聚類結果，單獨進行關注和轉發關係聚類時，大部分的用戶會劃歸到兩類。第一類為財經類和時政類資訊源，第二類為明星、娛樂和時尚生活類資訊源。劃歸到另外兩類的使用者數目較少，只有1–2個用戶。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

在6,737個用戶的較大規模樣本中，進行聚類分析，驗證方法的有效性。對聚類指標集矩陣Z_g，採用「內部指標」評估聚類效果，具體DB指數值參見表六。

表六　用戶譜聚類方法性能DB指數值比較表

	僅使用關注關係	僅使用轉發關係	並集處理	交集處理	二級處理	混合二級處理
DBI	0.003	0.002	1.53	0.38	0.27	0.18
群體數量	3,687/2,968/61/21	4,860/1,828/34/15	4,950/1,146/457/184	3,893/2,632/485/127	2,693/2,133/1,292/619	3,596/2,049/694/398

如表六所示，僅使用關注關係和轉發關係的DBI值最低，理論上性能最好，但觀察聚類結果，發現這種聚類性能的提升，實際是以劃分不平衡為代價的。微博用戶被主要劃歸到時政和娛樂兩類群體，另外兩類群體數目非常少，只有10–60個。結合DB指數值和聚類群體規模來看，混合二級仍然具有性能優勢，是可以進行實際推廣的聚類方法。

真實的微博用戶中，存在機器人使用者、純資訊轉發使用者等異常情況。微博用戶聚類，可以有效識別機器人用戶、重複發帖使用者等情況。對於真實聚類結構為m的情況，$k = m + 2$就可以得到較好的聚類效果。但對原有的二級聚類法群體結構改變不大。再繼續增加k值，只會在原有的聚類群體中不斷分化出零散邊緣個體，但不會影響聚類的總體結構。

研究結論與未來發展

本文提出了一種基於用戶認知差異的微博用戶聚類方法。該方法不依賴微博用戶屬性和文本內容，而是根據微博用戶認知反映到媒介選擇和信息處理上的不同，完成用戶群體的劃分。實例證明，該微博用戶的聚類方法是可行和有效的，且具備較好的穩定性。

作為社交媒體用戶群體的劃分工具，該方法建立的關注、轉發二級資訊傳播聚類機制，能夠有效發現不同認知傾向的社團組織。這種微博用戶聚類方法，將龐雜的社交網絡中的受眾與資訊源，統一提取

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

到資訊傳播維度中進行分析，可以更為全面和準確地理解資訊在網絡中的傳播機制。

本研究採取混合二級聚類方法機制，將關注信息源和轉發信息源的並集作為一級聚類數據源，將關注信息源和轉發信息源的交集作為二級聚類數據源。這種聚類方法，由於數據利用較為充分，且兼顧了兩種不同認知和傳播行為，能夠更好描述用戶的資訊使用行為，在性能上表現最好。

目前對大規模的微博用戶進行聚類，從聚類結果來看，微博用戶群體基本保持了關注時政和娛樂的二元結構。由於二元結構的用戶體量太大，直接增加聚類數目，也難以實現用戶群體的細顆粒度的劃分。因此，聚類方法可在初步二元聚類的基礎上，再針對每個群體各自進行細化，能夠較好地實現基於認知屬性的用戶劃分。

下一步的研究，考慮從數據樣本的代表性、聚類方法的準確性和算法的效率等方面，對聚類方法進行改進。目前研究中的用戶樣本的選取，盡可能考慮了抽取樣本的均衡性，但聚類方法仍需在更大規模用戶中得到檢驗。未來的研究可考慮將微博用戶進行分層抽樣，檢驗聚類方法的有效性。現有的聚類方法還面臨數據遺漏和噪聲數據干擾的挑戰：數據受到樣本抓取環境的限制，新浪微博上顯示的關注和評論僅能顯示前20或50頁的內容。當用戶關注數目或評論數目超過這一上限，會導致部分數據遺漏的問題；實際微博環境中，在用戶不知情的情況下，關注和轉發的信息源中經常會出現機器人用戶或者新浪微博自行添加的博主。這些虛假信息的噪聲數據，會影響聚類的效果。未來的研究可考慮在聯合譜聚類的處理基礎上，結合微博用戶的個人認證信息、傳播行為、微博文本，進一步提高聚類準確度。目前的聚類方法由於涉及到矩陣的乘法運算，計算複雜度較高，而隨著聚類用戶數量的增加，用戶關注和轉發的資訊源數量會成指數級地增長，因此聚類方法的效率還面臨大樣本的挑戰。聚類方法的改進，將考慮使用並行算法的大數據技術，提高算法的效率。

未來研究可以考慮將這種聚類方法應用到網絡議程設置的研究領域。2011年，郭蕾和麥庫姆斯(McCombs)借鑒了網絡科學的理論框

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

架，提出了網絡議程設置理論，以應對新媒體傳播環境對經典的議程設置理論的挑戰（Guo & McCombs, 2011）。在新媒體環境下，媒介難以準確掌握受眾的基本情況，進行用戶分析，評估傳播效果。論文提出的微博用戶聚類方法在對用戶聚類的同時，對用戶群體關注、轉發的資訊源也進行了劃分，為網絡議程設置理論的實證分析，構造了具體的分析環境。

「湯蘭蘭事件」話題的傳播時間跨度長，與公共話題通常7–10天的傳播週期相比，「湯蘭蘭事件」在社交媒體的傳播歷經三個階段，時間跨度長達六個月。在對微博用戶完成聚類後，後續研究將納入時間維度，在動態資訊環境下，觀察用戶群體結構的演化情況，獲取社交網絡的動力學特徵。後續研究可考慮在網絡議程設置研究領域應用該方法，分析受眾的認知和觀點是否發生了改變，會受到哪些因素的影響，評估議程設置效果。

當前的聚類方法是通過抓取的微博用戶數據，來推測用戶的認知。未來的研究可考慮在大數據分析的基礎上，結合問卷調查、訪談等社會調查的方式，進行大小數據的混合分析，在個體層面對用戶認知和動機進行合理的解釋，指導用戶聚類方法的改進。

參考文獻

中文部分（Chinese Section）

丁緒武、吳忠、夏志傑（2014）。〈社會媒體中情緒因素對用戶轉發行為影響的實證研究 ——以新浪微博為例〉。《現代情報》，第11期，頁147–155。

Ding Xuwu, Wu Zhong, Xia Zhijie (2014). Shehui meiti zhong qingxu yinsu dui yonghu zhuanfa xingwei yingxiang de shizheng yanjiu—Yi Xinlang weibo weili. *Xiandai qingbao*, *11*, 147–155.

中國互聯網絡信息中心（2017年1月11日）。〈2016年中國互聯網新聞市場研究報告〉。取自中國互聯網絡信息中心官網，http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/mtbg/201701/P020170112309068736023.pdf。

Zhongguo hulian wangluo xinxi zhongxin (2017, January 11). 2016 nian Zhongguo hulianwang xinwen shichang yanjiu baogao. Zhongguo hulian wangluo xinxi

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

zhongxin guanwang. Retrieved from http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/mtbg/201701/P020170112309068736023.pdf.

田占偉、王亮、劉臣(2015)。〈基於複雜網絡的微博信息傳播機理分析與模型構建〉。《情報科學》,第9期,頁15–21。

Tian Zhanwei, Wang Liang, Liu Chen (2015). Jiyu fuza wangluo de weibo xinxi chuanbo jili fenxi yu moxing goujian. *Qingbao kexue*, *9*, 15–21.

任薇、邱玉輝(2014)。〈基於微博網絡的社會公共事件輿論傳播研究〉。《西南大學學報(自然科學版)》,第6期,頁195–200。

Ren Wei, Qiu Yuhui (2014). Jiyu weibo wangluo de shehui gonggong shijian yulun chuanbo yanjiu. *Xinan daxue xuebao (ziran kexue ban)*, *6*, 195–200.

安璐、周亦文(2020)。〈恐怖事件情境下微博資訊與評論用戶的畫像及比較〉。《情報科學》,第4期,頁9–16。

An Lu, Zhou Yiwen (2020). Kongbu shijian qingjing xia weibo zixun yu pinglun yonghu de huaxiang ji bijiao. *Qingbao kexue*, *4*, 9–16.

李運華、汪祖柱、葉燕霞、張星星(2016)。〈基於熱點事件的微博用戶行為聚類實證分析〉。《情報探索》,第2期,頁75–79、83。

Li Yunhua, Wang Zuzhu, Ye Yanxia, Zhang Xingxing (2016). Jiyu redian shijian de weibo yonghu xingwei julei shizheng fenxi. *Qingbao tansuo*, *2*, 75–79, 83.

林燕霞、謝湘生(2018)。〈基於社會認同理論的微博群體用戶畫像〉。《情報理論與實踐》,第3期,頁142–148。

Lin Yanxia, Xie Xiangsheng (2018). Jiyu shehui rentong lilun de weibo qunti yonghu huaxiang. *Qingbao lilun yu shijian*, *3*, 142–148.

知微數據(2018年1月30日)。〈湯蘭蘭性侵案〉。取自知微數據官網,https://ef.zhiweidata.com/event/46dbb7a74a6659bb10002006/profileV2。

Zhiwei shuju (2018, January 30). Tang Lanlan xingqin an. Zhiwei shuju guanwang. Retrieved from https://ef.zhiweidata.com/event/46dbb7a74a6659bb10002006/profileV2.

苑衛國、劉雲、程軍軍、熊菲(2013)。〈微博雙向「關注」網絡節點中心性及傳播影響力的分析〉。《物理學報》,第3期,頁502–511。

Yuan Weiguo, Liu Yun, Cheng Junjun, Xiong Fei (2013). Weibo shuang xiang "guanzhu" wangluo jiedian zhongxinxing ji chuanbo yingxiangli de fenxi. *Wuli xuebao*, *3*, 502–511.

查魯·艾格華(2016)。《社會網絡數據分析》(陳哲譯)。武漢:武漢大學出版社。(原書Charu C. Aggarwal [2011]. *Social network data analytics.* Boston, MA: Springer.)

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

Chalu Aigehua (2016). *Shehui wangluo shuju fenxi* (Chen Zhe, Trans.). Wuhan: Wuhan daxue chubanshe. (Original book: Charu C. Aggarwal [2011]. *Social network data analytics.* Boston, MA: Springer.)

姚彬修、倪建成、于蘋蘋、李淋淋、曹博（2017）。〈基於多源資訊相似度的微博使用者推薦算法〉。《計算機應用》，第5期，頁1382–1386。

Yao Binxiu, Ni Jiancheng, Yu Pingping, Li Linlin, Cao Bo (2017). Jiyu duoyuan zixun xiangsidu de weibo shiyongzhe tuijian suanfa. *Jisuanji yingyong*, *5*, 1382–1386.

袁園、孫霄凌、朱慶華（2012）。〈微博用戶關注興趣的社會網絡分析〉。《現代圖書情報技術》，第2期，頁68–75。

Yuan Yuan, Sun Xiaoling, Zhu Qinghua (2012). Weibo yonghu guanzhu xingqu de shehui wangluo fenxi. *Xiandai tushu qingbao jishu, 2*, 68–75.

唐曉波、羅穎利（2017）。〈融入情感差異和用戶興趣的微博轉發預測〉。《圖書情報工作》，第9期，頁102–110。

Tang Xiaobo, Luo Yingli (2017). Rongru qinggan chayi he yonghu xingqu de weibo zhuanfa yuce. *Tushu qingbao gongzuo, 9,* 102–110.

席運江、杜蝶蝶、廖曉、仉學紅（2020）。〈基於超網絡的企業微博用戶聚類研究及特徵分析〉。《數據分析與知識發現》，第8期，頁107–118。

Xi Yunjiang, Du Diedie, Liao Xiao, Zhang Xuehong (2020). Jiyu chao wangluo de qiye weibo yonghu julei yanjiu ji tezheng fenxi. *Shuju fenxi yu zhishi faxian*, *8*, 107–118.

陳姝、竇永香、張青傑（2017）。〈基於理性行為理論的微博用戶轉發行為影響因素研究〉。《情報雜誌》，第11期，頁147–152、160。

Chen Shu, Dou Yongxiang, Zhang Qingjie (2017). Jiyu lixing xingwei lilun de weibo yonghu zhuanfa xingwei yingxiang yinsu yanjiu. *Qingbao zazhi*, *11*, 147–152, 160.

張春華（2012）。《網絡輿情的社會學的闡釋》。北京：社會科學文獻出版社。

Zhang Chunhua (2012). *Wangluo yuqing de shehuixue de chanshi*. Beijing: Shehui kexue wenxian chubanshe.

張翼、李培林、陳光金（2013）。《社會藍皮書：2014年中國社會形勢分析與預測》。北京：社會科學文獻出版社。

Zhang Yi, Li Peilin, Chen Guangjin (2013). *Shehui lanpishu: 2014 nian Zhongguo shehui xingshi fenxi yu yuce*. Beijing: Shehui kexue wenxian chubanshe.

許小可、胡海波、張倫、王成軍（2015）。《社交網絡上的計算傳播學》。北京：高等教育出版社。

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

Xu Xiaoke, Hu Haibo, Zhang Lun, Wang Chengjun (2015). *Shejiao wangluo shang de jisuan chuanboxue*. Beijing: Gaodeng jiaoyu chubanshe.

黃敏（2020）。〈議題關聯與關係建構——《紐約時報》有關中國扶貧報導的媒體網絡議程分析〉。《新聞與傳播研究》，第3期，頁21–36、126。

Huang Min (2020). Yiti guanlian yu guanxi jiangou—*Niuyue shibao* youguan Zhongguo fupin baodao de meiti wangluo yicheng fenxi. *Xinwen yu chuanbo yanjiu*, *3*, 21–36, 126.

新浪微博數據中心（2021年3月12日）。〈2020微博用戶發展報告〉。取自新浪微博數據中心官網，https://data.weibo.com/report/reportDetail?id=456。

Xinlang weibo shuju zhongxin (2021, March 12). 2020 weibo yonghu fazhan baogao. Xinlang weibo shuju zhongxin guanwang. Retrieved from https://data.weibo.com/report/reportDetail?id=456.

熊回香、蔣武軒（2017）。〈基於標籤與關係網絡的用戶聚類推薦研究〉。《資料分析與知識發現》，第6期，頁36–46。

Xiong Huixiang, Jiang Wuxuan (2017). Jiyu biaoqian yu guanxi wangluo de yonghu julei tuijian yanjiu. *Ziliao fenxi yu zhishi faxian*, *6*, 36–46.

劉繼、李磊（2013）。〈基於微博用戶轉發行為的輿情信息傳播模式分析〉。《情報雜誌》，第7期，頁74–77、63。

Liu Ji, Li Lei (2013). Jiyu weibo yonghu zhuanfa xingwei de yuqing xinxi chuanbo moshi fenxi. *Qingbao zazhi*, *7*, 74–77, 63.

英文部分（English Section）

Abdullah, N. A., Nishioka, D., Tanaka, Y., & Murayama, Y. (2017, January). *Why I retweet? Exploring user's perspective on decision-making of information spreading during disasters*. Paper presented at 2010 International System Sciences Annual Conference, Hawaii.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*(4), 644–675.

Boyd, D., Golder, S., & Lotan, G. (2010, January). *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter*. Paper presented at 2010 International System Sciences Annual Conference, Hawaii.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646.

Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The geospatial characteristics of a social movement communication network. *PloS One*, *8*(3), e55957.

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

Gomez, S., Diaz-Guilera, A., Gomez-Gardenes, J., Perez-Vicente, C. J., Moreno, Y., & Arenas, A. (2013). Diffusion dynamics on multiplex networks. *Physical Review Letters, 110*(2), 028701–028705.

Guo, L., & McCombs, M. (2011, August). *Toward the third level of agenda setting theory: A network agenda setting model.* Paper presented at the Association for Education in Journalism & Mass Communication, St. Louis.

Hill, A. L., Rand, D. G., Nowak, M. A., & Christakis, N. A. (2010). Emotions as infectious diseases in a large social network: The SISa model. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1701), 3827–3835.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 154–174.

Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, *36*(4), 593–617.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2), 561–569.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Li, B., Dittmore, S. W., Scott, O. K., Lo, W. J., & Stokowski, S. (2019). Why we follow: Examining motivational differences in following sport organizations on Twitter and Weibo. *Sport Management Review*, *22*(3), 335–347.

McGregor, S. C., & Vargo, C. J. (2017). Election-related talk and agenda-setting effects on Twitter: A big data analysis of salience transfer at different levels of user participation. *The Agenda Setting Journal*, *1*(1), 44–62.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Peng, H. K., Zhu, J., Piao, D., Yan, R., & Zhang, Y. (2011, December). *Retweet modeling using conditional random fields*. Paper presented at 2011 International Conference on Data Mining Workshops, Vancouver.

Petrovic, S., Osborne, M., & Lavrenko, V. (2011, July). *Rt to win! Predicting message propagation in Twitter*. Paper presented at 2011 International AAAI Conference on Web and Social Media, Barcelona.

Phelan, O., McCarthy, K., Bennett, M., & Smyth, B. (2011, April). *Terms of a feather: Content-based news recommendation and discovery using Twitter.* Paper presented at European Conference on Information Retrieval, Berlin.

Poell, T., De Kloet, J., & Zeng, G. (2014). Will the real Weibo please stand up? Chinese online contention and actor-network theory. *Chinese Journal of Communication*, *7*(1), 1–18.

Copyrighted material of: School of Journalism and Communication, The Chinese University of Hong Kong;
School of Communication, Hong Kong Baptist University (2022). Published by The Chinese University of Hong Kong Press.
ALL RIGHTS RESERVED.

Quercia, D., Capra, L., & Crowcroft, J. (2012, May). *The social world of Twitter: Topics, geography, and emotions*. Paper presented at 2012 International AAAI Conference on Web and Social Media, Dublin.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). *Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network.* Paper presented at 2010 International Conference on Social Computing, Minneapolis.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., & Wu, T. (2009, March). *Rankclus: Integrating clustering with ranking for heterogeneous information network analysis.* Paper presented at 2009 International Conference on Extending Database Technology: Advances in Database Technology, Shanghai.

Tang, L., & Liu, H. (2010). Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *2*(1), 1–137.

Trivedi, S. K., & Dey, S. (2016, March). *A comparative study of various supervised feature selection methods for spam classification.* Paper presented at 2016 International Conference on Information and Communication Technology for Competitive Strategies, Surabaya.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Xu, Z., Ru, L., Xiang, L., & Yang, Q. (2011, August). *Discovering user interest on Twitter with a modified author-topic model*. Paper presented at 2011 International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon.

Yun, J. T., Pamuksuz, U., & Duff, B. R. (2019). Are we who we follow? Computationally analyzing human personality and brand following on Twitter. *International Journal of Advertising*, *38*(5), 776–795.

Zhao, X., Sala, A., Wilson, C., Wang, X., Gaito, S., Zheng, H., & Zhao, B. Y. (2012, November). *Multi-scale dynamics in a massive online social network.* Paper presented at Internet Measurement Conference, Boston.

本文引用格式

周勝(2022)。〈基於認知的微博用戶聚類方法研究——以「湯蘭蘭事件」為例〉。《傳播與社會學刊》，第59期，頁177–209。