中文摘要關鍵字：史料資訊探勘、社會網路分析、資訊檢索

Abstract

We proposed a novel historical exploration method based on lexical co-occurrence network, information retrieval (IR) technic and social network (SNA) analysis metrics analysis. Plus, based on our previous research result, an improved version of Anchor-n-Gram (ANG), Compound Term Aggregation (CTA) was proposed, this algorithm will help us find out the most suitable form of language representation instances based on language-dependent sentence structure and statistical-based meaning evaluation algorithm TF-IDF. All these methods combined can tackle tremendous volume of text-based material and could help people not only explore from the detail discourse but also discover the profound understandings or emergent information from the aggregated results. In the specified dataset and research results, we found that the "228 Incident - Taiwan News Archive" reveals huge difference and significant consistency among our three underlining news source groups : "Government" , "People" and "Government with inclination of People". Especially on these perspectives: "Society Atmosphere", "Sentiments and Emotions", "Public Safety", "Armed Suppression", "Citizen Outrage" and "Outbreak".

Keywords: historical text mining, social network analysis, information retrieval, critical discourse analysis.

作者英文名及英文單位

Ning, Ke-Chih[1]、Liu, Jyi-shane[1]、Hsueh, Hua-Yuan [2]、Tsai, Ming-Feng [1]
[1] Department of Computer Science, NCCU
[2] Graduate Institute of Taiwan History , NCCU

以概念場域相似度為基礎之論述輪廓與情感分析
– 以二二八事件臺灣本地新聞史料為例

Discourse Profile and Sentiment Analysis by Concept Domain Similarity - A Case Study of the "228 Incident - Taiwan News Archive"

甯格致[1]、劉吉軒[1]、薛化元[2]、蔡銘峰[1]
[1]國立政治大學資訊科學系
[2]國立政治大學台灣史研究所

數位人文之資訊探勘

數位人文的初期研究多半屬大量史料文本數位化的工作，或是藉由領域專家制定產生權威辭庫，此類型資訊品質雖高，但主觀判斷及應用情境相依程度亦較高，對於方法流程之轉化再應用難度較高。文史資料應用資訊方法於提取資訊仍有許多可改善的空間，像是產生資訊過於零碎、意義過於廣泛、或是難以被解釋等問題。人們更感興趣的應是以詞彙形象呈現的共通議題及其之間的關係樣貌，而非單面向資訊累計的結果，因此在當前大量數位資料的時代裡，文史學者的精神若能與資訊方法更細緻地融接，脫離既有文史精讀及資訊檢索框架的束縛，利用自然語言呈現的特性，發展還原語境中語意之詞彙方法，並以其為基礎進一步揭露不同層次概念相互之間關係的結構，發展合適的系統化資訊探勘方法，應有助於數位人文下一階段的發展。

以人事物所構成的概念框架為索引從語境中提取重要資訊

文本中的人、事、物、情感概念，可藉由概念及其關係之堆疊凝聚成更具體的概念群，構成一概念網路。文史領域專家的知識可以提供我們應具備的概念框架以及所對應到的可能字詞樣貌。例如當我們定義概念框架為以下結構：總概念→子概念→議題→構面，我們可藉由文本資料建立出：總概念(二二八事件(→子概念{事件爆發}→議題{查緝}→構面{查緝員、傷亡}，類似如此的概念架構。此外我們仍須考量：1)背景語言知識；2)詞彙共現關係；3)複合詞結構。

具意義敏感度之詞彙產生方法 Meaning-Sensitive Compound Term Aggregation

目標議題導引關聯式網路模型(Anchor-n-Gram, ANG)是以指定概念詞彙為核心進行資訊擷取建立詞彙關聯網路，並提出詞彙合併動態規劃演算法，解決斷

詞的結果往往會將原本較具有意義的詞彙分解成基本的詞性(Part-Of-Speech, POS)，使得詞彙呈現的意義太廣泛而失去原始語境所反映概念的問題(2014, 甯格致等)。ANG 法將詞性單元的詞彙重新整併為意義較完整的詞彙，但若單以複合詞結構原則進行合併，將有可能發生詞彙過長，詞頻過低，導致在權重篩選中被捨棄的狀況，失去與其他相關語境關聯的可能性。

本研究提出 ANG 法的改良版 Compound Term Aggregation (CTA)，依據文本樣貌計算詞彙意義含量並配合詞頻決定最佳合併結果之評估法：Meaning-Sensitive CTA (MS-CTA)，其運作架構為：

1) CTA 階段：依據複合詞結構規則將詞彙合併至收斂。

2) CTA-based N-gram 階段：依據 CTA 階段建立的詞彙及其詞性，進行 POS-based N-gram，N=1~K，其中 K 為該詞彙所包含的詞性總數。

3) CTA-TF-IDF 階段：於候選詞彙中逐詞計算詞頻及文本出現次數，並以 double normalization 及 Inverse Frequency Smooth 方法平滑計算 TF-IDF 數值，避免文本大小不一產生的偏差情況。

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

$$\mathrm{idf}(t,D) = \log(1 + \frac{N}{n_t})$$

4) CTA-Rank 階段：最後依據綜合規則之門檻挑選出優先且合適的詞彙(例如：TF-IDF 較高者、文本出現次數大於 1 者、不為單一助詞結構者、不為單一副詞結構者、以複合詞長者優先等規則)，完成 MS-CTA 計算程序。

關聯詞彙共現網路與概念場域相似度計算

MS-CTA 方法可有效地提取出較具意義且適合於建立關聯之詞彙，計算各字詞間距離之加權可獲得共現的關聯強度，構成基本雙邊結構關係 (dyad)，逐步施作即可擴展為三邊結構(triad)及群聚結構(clique)，再利用合適的社會網路結構指標進行計算，本研究採用自我中心網路(Ego Network)，以議題為核心，於概念網路中提取出相關聯的議題構面集合。

$$C_D^w(i) = \sum_j^N w_{ij}$$

各議題所建立出無向性具權重的構面關聯詞彙共現網路，可進一步再利用權

重中心性指標(Weighted Degree Centrality)，得到議題於各構面的關聯強度，強度越高則代表著該議題詞彙出現的次數越多、與其他構面伴隨出現的次數亦多、兩詞的意義或關聯性越相近。最後，我們應用餘弦相似度 (Cosine Similarity)，將向量及維度值代換為議題及所屬構面之關聯強度，進行相似度計算，可進一步呈現議題、子概念、總概念以及文本間的相似程度。綜合言之，本研究所提各方法其重要運作程序如下：

1) 由文史領域專家藉由欲探討的主題逐層定義出其總概念、概念及議題，構成一整體概念框架。
2) 使用 MS-CTA 方法逐各文本中計算擷取出較具有意義的詞彙。
3) 依據這些詞彙以及文本中的詞彙距離建立起詞彙關聯網路。
4) 藉由自我中心網路方法以給定之議題為中心點，計算權重中心性指標得到屬於該議題下之無向性具權重的構面詞彙關聯網路。
5) 提取出各構面詞彙關聯網路中的構面及其與議題之間的關聯強度值，依據向量及維度的形式轉換成為各種不同之議題向量，並依據餘弦相似度方法計算出議題向量之間的相似度。
6) 依循概念框架進行平行及垂直之綜合(平均)比較結果。

二二八事件臺灣本地新聞史料彙編

”二二八事件臺灣本地新聞史料彙編”資料共分為十大報刊(設立屬性屬於官方者有三刊、屬民間者有六刊、屬官偏民者一刊)，本研究篩選民國 38 年 3 月 13 日查封日為止之文字共 22 萬餘字。在文史專家的協力下從中挑選建立出概念框架的內容：

主概念：二二八事件

子概念：

事件爆發 (議題：查緝、私煙、毆打、死傷、市內、市民)；

武力制壓 (議題：警察、軍隊、憲兵、軍警、戒嚴) ；

引起公憤 (議題：不幸、責任、傷亡、受傷、嚴辦、撫卹、專賣局)；

社會氛圍 (議題：治安、人民、民眾、學生、民主、糧食、失業、秩序)。

並以前段所述關聯詞彙共現網路及概念場域計算方法應用於文本中，進行四

個概念於三種報刊屬性(官方、民間、官偏民)之論述輪廓分析。在情感分析方面，以情感語料庫(2007, Chen et al.)對議題向量中呈現之情感詞彙進行極性的偵測及加權，計算出四個概念於三種屬性報刊之正負向情感分布，再檢定其共變異性。

分析及結論

在論述輪廓相似度分析中，”社會氛圍”的各報平均相似度最高(32%)，又以”中華日報”(官方)與”和平日報”(官偏民)的相似最高(62%)，”中華日報”(官方)與”中外日報”(民間)次之(55%)，”和平日報”(官偏民)與”中外日報”(民間)再次之(49%)。”引起公憤”的平均相似度最低(12%)，”和平日報”(官偏民)與”民報”(民間)以及”和平日報”(官偏民)與”興台日報”(民間)之相似度為最低(9%)。綜合各子概念及各屬性，”官偏民”與”官方”的相似度為最高，再來是”官方”與”民間”，相似度最低為”官偏民”與”民間”。

在情感極性的分析中，我們發現”民間”情感幾乎皆傾向於負向，”官方”及”官偏民”則傾向於正向，”武力制壓”最屬負面(62%)，”事件爆發”(57%)次之，”引起公憤”(55%)再次之；"民間”展現出最負向情感的”武力制壓”(62%)中反而呈現正向情感者為”官偏民”(63%)及”官方”(56%)。皆展現正向情感的”社會氛圍”中”官偏民”(76%)、”民間”(64%)及”官方”(57%)。在情感共變異報組分析中，正負向皆一致達顯著，”官方” 有 2 組，”官偏民”與”民間”有 3 組，”民間”有 4 組，而”官方”與”民間”及”官方”與”官偏民”則全無顯著。

其他亦有值得深入探討的現象，概略描述如下：

1. 各子概念中，“社會氛圍”在各報刊所呈現論述輪廓的相似度是最高的，在情感的用語上也是較偏為正向，審視文本及計算過程可看出人們對當時社會氛圍的呼籲、訴求及期望。

2. 屬”官偏民”之”和平日報”，在四個子概念的論述輪廓上其實仍是比較偏向”官方”的，但在情感的表達上則是比較傾向”民間”的。

3. 屬民間最大報的”民報”在情感面的描述上，有明顯偏向負向的情況。而屬於官方最大報的兩報 " 台灣新生報”及”中華日報”皆明顯傾向於正向的描述，僅”中華日報”於”治安”以及”人民”兩議題上，展現了特異強烈的負向論述。

4. “長官公署所屬機構與警備總部公報” 可能因為公報性質，其所展現的論述輪廓及情感與其他九份報刊有非常顯著的不同，因此在概念框架中的各議題符合程度相當低，情感詞彙的使用亦相當低，造成拉低了各項平均比較的相似度數值，然而相對比較之排名是不受影響的。

本研究提出具意義敏感度之詞彙產生方法 MS-CTA，有效地解決了一般使詞頻統計方法中浮現之詞彙意義過於廣泛或不足的常見問題，並以意義敏感性偵測方式，從文本中提取出較具有意義的詞彙，並結合社會網路分析模型及餘弦相似度的運用，在統計的逐步浮現及驗證之下，從目標文本中大量且快速有效地發現了許多具有潛在豐富意涵的資訊，相信可作為後續相關研究的參酌運用。