《人文及社會科學集刊》
第二十七卷第四期（104/12）, pp. 683-718

社會科學研究中的文字探勘應用：以文意爲基礎的文件分類及其問題*

陳世榮
中國文化大學行政管理學系副教授

隨著電子典藏技術的精進，文字探勘技術逐漸受到重視，本文以社會科學研究在文意區別上的需求，評估監督式機器學習對非結構、複雜文本的分類效果，並就所見問題提出分析與建議。本文從文字探勘與內容分析文意區別上的差異與共通性出發，繼而以新聞報導爲分析資料，針就特定文件意向，遵循一般文字探勘程序，以支持向量機與簡易貝式分類器執行文件分類評估。分析結果指出，文字探勘對於複雜文意的判讀效果值得肯定，但經由共詞網絡分析也發現，文件的編撰風格將影響文件分類的效果。建議研究者在資料處理初期，應反覆評估研究目的、資料特性與分類器模型間的契合度。

關鍵字：文字探勘、文意區別、文件分類、機器學習、共詞網絡分析

壹、文字探勘應用及其疑慮

隨著網路資訊發達，數位典藏盛行，社會科學研究史無前例地受惠於電子資料的應用，類型多樣的資料延伸了知識探求的觸角，歷時性的檔案文本開啓了深究社會結構與行爲趨勢的機會。然而，大數據（big data）的利用，也意味著社科研究在資料處理能力上的挑戰，傳統所依賴的研究方法，諸如內容分析法，必須做相應的調整。所幸，伴隨著資訊科學的發展，資料解析

＊作者感謝二位匿名審查人的寶貴意見。
收稿日期：103 年 9 月 11 日；接受刊登日期：104 年 7 月 28 日

與處理一直受到高度關切，並已發展成體系龐大且複雜的知識領域，這無疑提供社會科學界在未來數位資料應用上極爲重要的支持與推力。

應用層面上，過去二十年間，在商業利益的推波助瀾之下，對電子化訊息解讀的需求陡升，促使資料探勘（data mining）熱潮，進一步延伸出文字探勘（text mining）的應用。所謂「文字探勘」指的是，就所蒐集的特定巨量文件，執行編輯、組織與分析的過程，以發現其間隱含的特徵關聯或新穎有趣的模式，提供分析師或決策者特定的訊息（Blake, 2011: 125–126; Sullivan, 2001: 326）。由於探勘過程高度仰賴電腦計算與模型運用，得以處理巨量資料，使得「自然語言處理」（natural language processing）技術獲得跨領域的應用價值，社會科學研究也不例外。

但是，這份期待也不是沒有限制的。首先，基於網際網路所引發的資訊爆炸與訊息超載，快速簡潔地萃取有意義及有價值的資訊，正是「自然語言處理」發展的目的。如同當前資訊科學所積極推動的語意網（Semantic Web）與人工智慧，其終極目標在建立一個計算機能閱讀的知識本體（knowledge ontology），使得電子媒介或網頁可以迅速與精確地傳達所需的知識（戚玉樑、蔡明宏，2007；黃居仁等，2004）。這當然是一項浩大的工程，目前計算語言學相關學理與技術仍在演進階段，與前述終極目標仍有相當距離。如此，對於社科研究者而言，最大困擾在於缺乏一個完整權威的介面，以確保所處理資料能獲得領域內外的認同，這種情況在繁體中文世界尤其如此。然而，一套可被廣爲接受的技術系統，需要經過長時間的沉澱，以及不同應用面的淬煉。如此說來，社會科學研究者對於文字探勘運用應該抱持積極開放的態度，使得社會科學的專業領域考量，也能在自然語言處理技術發展的過程中扮演一定角色。

社科研究在文字探勘應用所面臨的第二項困難在於跨領域的知識吸納。文字探勘技術背後是一個急速發展、體系龐大的資訊科學領域，除了原本具有資訊背景的少數人之外，對於大部分社科研究工作者而言，僅僅涉入基本的應用技術都將遭遇極大的挑戰，因此，跨領域合作會是一個好的因應策略（例如，林琬眞等，2012）。然而，眞正的問題還來自於不同領域對於新技術方法的疑慮，這種疑慮雖也有來自知識的隔閡，但更可能是因爲領域差異下

所涉及認識論、方法學、與問題設定的差異。這種疑慮無法完全藉由跨領域合作獲得解決，任何一套計算語言技術在不同領域的應用，可能被視爲是一種黑箱，甚至被貶爲過度炫技卻不能解決個別領域所關心的課題或符合方法學要求。因此，如何從應用的角度反向檢視文字探勘的方法與效果，是文字探勘技術推廣進程中相當重要的工作，而這也正是本文的目的所在。

爲此，本研究以新聞文本爲對象，應用文字探勘方法，解析文件潛在意義，進行文件分類，並以此實驗設計，評估文字探勘在區別文件意義上的可行性與限制。此處所謂文件潛在意義，指文件內含的意義定向，這種意義定向往往涉及訊息傳送與解讀者的社會系絡，不容易直接由詞語的外顯意義來區別，傳統上社科研究均以人工標記加以判讀，亦即採用內容分析法。以下依循內容分析與文字探勘的差異，說明本文的主要論旨。

貳、內容分析與文字探勘的比較

一、內容分析程序

語言與文字是人類社會獨有的文化遺產，透過語言與文字，社會的溝通、傳播、知識、文化才有可能，語言文字無疑是社會科學重要的研究對象。社會科學研究中早已奠下文本分析與內容分析的傳統，其中，對於詞語的計量表達早爲政治學者拉斯威爾（Harold D. Lasswell, 1965）所倡議。

由於社會科學對於文本有極爲多樣的分析進路，此處取與文字探勘操作程序最接近的內容分析法作爲考察比較的對象。不過，內容分析在社科研究中，也因學門與課題不同而有極爲不同的應用形式，此處無法做系統性比較，更無意評論究竟何種文本分析途徑適合社會科學研究。[1] 這裡主要是透過比較，指出內容分析與文字探勘的異同，進而凸顯社會科學研究所關心的議題。爲求聚焦，以下選擇 Krippendorff 的定義與分類，作爲討論的起點。

Krippendorff 將內容分析定義爲一種從文本推論出其使用系絡（contex of use），並能重製且具效力的技術（Krippendorff, 2013: 24）。進一步，Krip-

1 有興趣的讀者可參閱其他相關專論（例如，游美惠，2000；瞿海源，1982；蘇中信，2012 等）。

pendorff從實際應用做出三種分類，亦即，文本驅動（text-driven）、問題驅動（problem-driven）以及方法驅動（method-driven）等三種分析（Krippendorff, 2013: 355-372）。顧名思義，方法驅動是基於所使用方法而設計的研究，文本驅動分析則是在確認某種文本與研究課題有關後，對該文本所進行的內容探索，換言之，此一途徑重視文本內涵的自我展現，例如紮根理論（grounded theory）（Glaser and Strauss, 1967）。然而，在多數的情況下，研究者對內容分析的使用是基於研究課題的需要，屬於問題驅動。Krippendorff對問題驅動分析有更爲詳盡的說明，並指出其程序爲：問題形成、確認資料、蒐集資料、定義及發掘分析單元、取樣、發展編碼分類與執行標記、選擇分析程序、採納分析標準等等（Krippendorff, 2013: 357-370）。

內容分析具有融合質化與量化分析，提供定量數據，以非接觸性研究減少主觀涉入等優點，頗受研究者歡迎。不過，上述程序中的單元定義、編碼標記、分析決策等，多仰賴人工判斷。雖然目前有軟體可資協助，但人力與時間成本仍高，若是面對巨量資料更是難以應付，因此社科研究者有借重文字探勘技術的需求。

二、文字探勘程序

文字探勘技術從圖書分類的使用，發展到對非結構文本的資訊萃取（information extraction），期間吸納了來自資料庫、資料檢索、人工智慧與機器學習、計算語言學、圖書資訊學的領域知識（Miner et al. eds., 2012: 3-27）。所謂非結構文本是指以日常生活中經常面對的，以自然語言所構成的文本，是未經特殊分類或人工關聯處理的自由文字（陳文華等，2003: 42；Miner et al. eds., 2012: 44）。文字探勘主要程序包含資料檢索與處理、斷詞（word segmentation）、特徵萃取（feature selection）、分類與集群、文本表示與詮釋等程序（曾元顯，2002；賴志遠等，2009: 19-21；Feldman and Sanger, 2007: 13-17; Miner et al. eds., 2012: 53-72），細部程序及其內在邏輯將於下文以實例詳述，但它在不同領域與不同實務上有不同的功能強調，而且不同功能的技術發展也不一致（Miner et al. eds., 2012: 30）。不過，從解析可用訊息此一目的來看，內容分析法與文字探勘確有其共通性。

文字探勘應用有幾點優勢。首先，其分析對象多半屬電子化文本，透過線上取得極爲方便，更重要的是，若具備即時更新功能，文字探勘可以處理動態資料；文字探勘對於龐大資料的處理能力，使得此前無法觀察的訊息與知識可以獲得論證；再次，由於可處理的資料形式多元豐富，且歷時穩定，可提供更完備的論證效力；而自動化技術不僅能克服主觀判斷的失誤，更能加快處理速度，大大節省時間與人力成本；最後，它有當前蓬勃發展的資訊學科支持，技術得以不斷精進，目前已有許多成熟的技術可適用於不同類型的文件分析，尤其是從非監督進展到監督式學習（supervised learning），可以就小量資料執行特徵萃取後，預測並描述巨量資料的意義，極具應用潛力。

簡略來說，文字探勘演算是以文件的特徵表現爲基礎，而非文件中的敘事內容，主要特徵包含字母、字、詞、概念等層次。就尋求文件意義價值來說，詞語與概念層次的效果會更好，這通常是藉由複雜的統計、語法，或者混合二者的分類器（classifier）來達成。分類器往往需要參照外部知識確認概念的意義與關聯，例如預先備妥的領域本體（domain ontology）或詞典（lexicon），社科研究偏向執行個人式專題研究，則多半仰賴有預先標記的訓練資料。在實作上，詞彙在去除原來語句與敘述結構後，以向量方式儲存成語料成爲分析對象。

這種把文本單元化、法則化進而汲取意義或訊息的方式，對向來重視論述與溝通的社會科學研究者而言很難不抱持疑慮，甚至不能接受。[2] 其中的理由很明白，因爲，社會事件參差消長，溝通語意繁雜多變，語境差異主客有別，更不用說，還有語用學所提出的「言外之意」以及「語言行爲」（speech act）等細微且複雜的表達。自然語言處理技術顯然有它的限制（Hand, 2006; Sullivan, 2001: 40），[3] 或者更正確地說，文字探勘在講求精確與效率之間存

2 Laver and Garry, 2000: 625；另 Bock 對於目前相當流行的文字雲（word cloud）批評指出，文字雲並不具備科學或推論意義，它僅是基於「文字頻率代表某種意義」的假設下，提供了讀者一種無限制的解讀（Bock, 2009）。

3 事實上，新一代技術已致力深層挖掘（deep discovery），並設法判別慣用語（idiom）、嘲諷（sarcasm）、諷刺（innuendo）等隱晦的語意（Miner et al. eds., 2012: 12），或如臺灣學者已著手建立的情緒分析（李政儒等，2012）。

在著兩難（Feldman and Sanger, 2007: 5）。不過，Feldman 與 Sanger（2007: 61）對此有頗爲中肯的討論，他們以語法剖析爲例指出，當前技術在文本解讀上仍無法兼顧效率與精確，對於巨量語料，使用當前常用的演算技術執行分析，仍有硬體與時間成本過高的問題，若採淺層剖析（shallow parsing），雖會遺漏部分訊息，但執行上將更爲快速、容易、穩健，因此在資訊萃取上，淺層分析仍有其實用性。俞士汶（2003: 317）更指出，以當前技術來說，「完全分析」不僅無法提升性能，有可能產生更多錯誤，二者在時間耗費上將出現極大的差距。

因此，在社科研究運用上也可以體認這種效率與精確的取捨，不需將深層解析的要求作爲檢視文字探勘效用的唯一基準，善用文字探勘解析多量文件的優勢。當然，這仍無法迴避其能否支持社會科學分析的問題。本研究主要聚焦在文件層次的意向分類，以回應社會科學研究的實務需求。

三、文件層次的意義區別

內容分析除了用於解析具明確意義的語彙之外，有時也用來發掘文本的內涵、意圖、效果，甚或言外之意（Krippendorff, 2013: 141）。文本解讀往往涉及訊息的傳送者、接送者、研究者以及相關的特定社會系絡，這便有別於從科學角度看待文本的方式。對於許多社科研究者而言，文本是由「人」去創作，並由「人」去解讀，本文本身並不具客觀性，同一文本不必然蘊含單一意義，不同的人也可以有不同的解讀，僅僅以字詞的表面意義作爲分析對象，窄化了文本的作用，而文本所傳遞、激發、觸動的效果，往往超越文字本身（Krippendorff, 2013: 24–31）。這種對文本的認識論立場不是所有社會科學途徑都會接受，但確實開展了文意解讀極爲活潑彈性的一面。不過在實際應用時，基於內容分析的方法學要求，社科研究者仍需以他自身的學術訓練，愼選與課題有關聯的文本，使文本的產製來源與訴諸對象能夠對應研究課題，並明白揭露推論設計與判讀程序，以建立可重複驗證以及合理有效的基礎。

在方法學上，文字探勘是在盡量減少人力判斷的條件下，借助各種演算法挖掘文本的意義。它的首要工作是將文本內容數據化爲詞彙出現頻率

（Miner et al. eds., 2012: 30），並以「文件—詞彙」分布形式，以向量空間呈現其初始資料（Blake, 2011: 127）。從社會研究者的觀點看，這種被稱爲詞袋（bag of words）的處理方式，直覺上是不可思議的，但相關經驗研究已證實它對於文本分類效果卻頗爲有效（Hopkins and King, 2010: 232; Pang et al., 2002: 83）。事實上，文字探勘原理與人工內容分析有異曲同工之妙。例如，人們在閱讀某篇報導時，可能在幾句之內便可以判讀該篇報導是否支持某特定意向，而不需從頭到尾細讀。相同的道理，文字探勘假設，只要能正確掌握少數關鍵的詞彙特徵，也可以區別文本的類屬。因此，儘管計算語言學將自然語言文本稱做「非結構」資料，但事實上，文字探勘假設文本存在著某種「結構」，只是這些結構鑲嵌在文本中，以致隱而不顯，必須加以挖掘（Feldman and Sanger, 2007: 3）。因此，若認定文字探勘只是種詞頻組合的計算，是過度簡化的誤解。在方法程序上，文字探勘不僅與內容分析有相通之處，且有低成本、處理巨量資料、運用電腦演算、減少人爲判斷等優點，可與內容分析取得互補效果。

在認識論上，文字探勘對所處理的文本資料，假設有一組詞語特徵可區別某一特定意義，因此也連帶假設存在一個「眞」的模型（Hopkins and King, 2010: 234）。對於文本中的多元意涵是以高維特徵空間加以區別（Cristianini and Shawe-Taylor, 2000: 26–49），探勘的目的就是找出與某一「眞實」意義匹配的概念描述。這與前述內容分析對文本解讀的觀點，存在著認識論上的差異。雖然在程序上內容分析也從詞彙區別意義，但考慮語言表述的多義與抽象使社科研究者多了一層認識論上的彈性，並設想文本解讀可能存在的不確定性。

基於以上分析，本研究主張，內容分析與文字探勘在意義區別上，固然有認識論上的差異，但這並不一定減損二者在方法學上的共通性，社科研究者可以在保有認識論的彈性下，於適當時機使用文字探勘進行文意區別或文件分類。若再考慮文本複雜度、傳播雜訊（noise）以及方法侷限，文字探勘與內容分析所得結果，只有彼此是否接近、能否互補的考量，而沒有何者可以反映唯一眞實的問題。基於這樣的觀點，社會科學研究的應用問題，便著落在實際使用的效果，亦即以文字探勘技術所獲致的文件分類，是否能趨

近人工判讀的結果，而這只能藉由經驗性評估來尋求解答。

基於此，本文以實驗設計方式，針對文件的意義區別，評估文字探勘技術對非結構、複雜文本的分類效果，這在文字探勘領域屬於文件分類的課題，同時也是社會科學領域經常需要應用的文件意義內容分析。考量社科研究的實際應用，除了文意區別效果的評估之外，本研究也觀察在使用監督式機器學習流程下，資料的檢索詞、關鍵詞數、文件數、類別大小，分類器以及資料編撰風格，對文字探勘的影響，從而提出在社會科學運用上所需注意的問題與改善策略。

參、研究設計與資料

一、研究設計

本研究希望從社會科學的應用角度，以特定議題爲對象，遵循文字探勘一般程序，驗證其文件意義分類的效果，並發現可能問題，以提供未來比較與改進的經驗依據。首先，本研究由線上檢索方式取得待分析之文件集，如此最貼近實際應用情境，亦即不採用計算語言學在驗證不同方法時所採用的測試檔本。所擷取的文件是從已知的文件脈絡中，以特定詞彙檢索出自然語言文本。依前述討論可知，除了檢索字詞所尋求的主題之外，文件集中的任何文件均存在K 個可能的其他概念或主題，爲確保文件內含足夠的複雜度，刻意選用一個較爲模糊的詞彙進行文件檢索。下文的分析中可以看出這種複雜度所可能衍生的問題與意義。

本研究選擇支持向量機（Support Vector Machine, SVM）與簡易貝式分類器（Naive Bayes Classifier）爲文件分類器，前者是廣獲好評的新型監督式學習分類器（Cristianini and Shawe-Taylor, 2000: 7; Pang et al., 2002: 83; Yang and Liu, 1999），後者已沿用多年，一直是探勘的基本工具，經常用來做參考比對，但有時效果極佳（Tufféry, 2011: 497）。監督式學習分類器是先以部分資料加以訓練，掌握目標意義的特徵，經建模與評估調校後，對所剩的大部文件進行預測，相較於單純的演算或統計分類，監督式訓練原則上將更貼近目標資料的特徵與結構，得以提升分類辨識的效能。

一般所謂監督式學習通常將手中資料區分爲三個獨立的部分，以訓練資料（training data）作爲機器訓練用，以受測資料（test data）作爲測試評估與建模之用，最後再對未經標記的目標資料（target data）進行預測。本研究以驗證評估爲主，資料僅區分訓練與受測資料。由於所檢索出的資料屬於自然語言，因此以傳統內容分析法先對所有資料進行人工分類標記。一方面，人工標記結果可作爲訓練資料，以提供機器學習之用，另一方面，人工標記結果納爲受測資料時，可供做文字探勘與內容分析的一致性檢驗。惟如前述，此處不將任何結果視爲眞切完滿的事實眞象，而是尋求二者的最小差異，以評估文字探勘在社科研究上的適切性。原則上，資料筆數愈大較能確保文件分類的效果與模型的精確性（Witten et al., 2011: 149-150），但在社科研究實務上往往面臨資料稀少，或基於研究需要，一次納入分析的資料量未必很大。爲此，本研究以時間區段檢索出較小量的文件集，藉此壓縮機器學習的空間，以瞭解文字探勘在小文件集的分析效果，由此可以觀察監督式分類器的穩定度與應用廣度。在經過效果評估後，本文將就所見問題做進一步分析推論，從而提出相關的注意事項與建議。

二、資料來源

資料來源方面，以社科研究常用的新聞電子資料庫爲對象，設定「公投」爲檢索詞，檢索自由時報與聯合報二報新聞文本，時間範圍取 2008 年總統大選前第一個月（2008.02.22～2008.03.21）、第三個月（2007.12.22～2008.01.21）以及第五個月的資料（2007.10.22～2007.11.21）（如表 1），合計檢索出 1556 篇報導。取樣的時間區隔，主要著眼於幾點研究設計的考量。首先，檢索資料在壓縮每一文件集的資料量，以提升探勘的難度。其次，在同一關鍵詞、相同時段的採集條件下，探勘不同的報紙報導，也可以觀察不同報紙對文字探勘的影響。最後，隨時間推移，相關事件本身會發生變化，因此可以觀察文字探勘處理不同時段文件效果的一致性。

上述資料內容是個已知但頗爲複雜的文本語境，以寬鬆的「公投」關鍵詞檢索的結果，所含報導內容、涉及事件，與表述方式的變異不小，恰可檢視文字探勘的效能。2008 年的立委與總統選舉決定著民進黨政權保衛與國民

黨期待的二次政黨輪替，選情異常緊繃，對於第七屆立法委員選舉，執政的民進黨拋出公投「討黨產案」，在野的國民黨則提出公投「反貪腐案」；而伴隨總統大選民進黨提出「以臺灣名義入聯」公投，國民黨爲求反制提出「以彈性名義返聯」公投，此期間各個公投案的正當性論述，經常融合著臺灣未來發展的大論述，與競選策略的枝節口水，媒體論述相當複雜。

本研究選擇單篇報導「是否支持民進黨或國民黨版公投」作爲分類依據，這是一種文意定向的分類。雖然，基於臺灣媒體特性，研究人員可以預知，自由時報支持民進黨版公投的報導偏多，而聯合報偏少，但不僅「公投」一詞本身已無意義區別作用，顯性詞意組合也很難作爲判讀依據，若加上可預期的特殊事件穿插以及「公投」一詞所涵蓋的不同公投案論述，對人工判讀與文字探勘都是一項考驗。無論採取何種途徑分析，文件分類的結果可提供正式研究問題的論證基礎，例如媒體報導是否平衡，或公投議題如何變化等，但這並非本文的關心議題。

以上的資料，以不同月份、不同報紙構成六組新聞報導文件集，並依據前述研究設計，先經過傳統內容分析進行文件標記，也就是以各篇文件「是否贊成民進黨版公投」加以分類（見附錄），再利用文字探勘技術進行文件分類，以比對人工判讀與機器學習分類的結果，此即文字探勘領域所稱的「精確率」，透過這樣的驗證得以評估文字探勘在社科分析上的適用性與可能遭遇的問題。

三、文意的內容分析

程序上，先以傳統內容分析方法對所獲資料進行分類，提供後續文字探勘比對與評估的基準。以「支持民進黨版或國民黨版公投」爲單一分析題目，爲求判讀與標記的明確性，以民進黨版公投爲基準建立一個三分類變數，編碼「支持」代表贊成民進黨版公投（等效於反對國民黨版公投），編碼「反對」代表反對民進黨版公投（等效於支持國民黨版公投），編碼「無關」代表中立、無關或無法判斷。判讀時，報導類型主要區分主觀意見報導、客觀事件報導，以及客觀事件報導中內含的反向主觀意見等三種（見附錄）。以三位大學畢業程度助理進行判讀，正式判讀前隨機選取 20 個文件，比對彼

此的判讀基準並做適切調正。文件歸類須至少有二人判讀相同，否則列屬無效分類。依據 Holsti 的信度計算方式，相互同意度（inter-judge agreement）等於二人共同認列數目乘以 2，再除以二人個別認列的總數，把兩兩的相互同意度加以平均即爲平均相互同意度 A（average inter-judge agreement），設 N 爲判讀人數，則複合信度 CR（composite reliability）可作爲內容分析的信度表示，公式記爲：$CR=(N\times A)/(1+(N-1)\times A)$（Holsti, 1969: 137）。

表 1 顯示判讀結果，六個文件集內容分析信度分別爲自由時報的 0.94、

表 1：資料檢索結果與人工分類判讀

	時間範圍	總筆數	編碼別	有效分類筆數	分析信度	無法認列數
自由時報	2007.10.22～2007.11.21	166	支持	99	0.94	37 (22.3%)
			反對	7		
			無關	23		
	2007.12.22～2008.01.21	315	支持	187	0.96	55 (17.5%)
			反對	12		
			無關	61		
	2008.02.22～2008.03.21	184	支持	117	0.96	33 (17.9%)
			反對	12		
			無關	22		
聯合報	2007.10.22～2007.11.21	224	支持	37	0.96	37 (16.5%)
			反對	112		
			無關	38		
	2007.12.22～2008.01.21	464	支持	45	0.94	115 (24.8%)
			反對	118		
			無關	186		
	2008.02.22～2008.03.21	203	支持	43	0.97	24 (11.8%)
			反對	66		
			無關	70		

資料來源：檢索並整理自自由時報，2007-2008；聯合報，2007-2008。

0.96、0.96，以及聯合報的 0.96、0.94、0.97，均超過一般內容分析的信度要求 0.80。不過，無法認列文件數仍占總筆數的 11.8% 至 24.8% 之間，這一方面顯示人工判讀的不穩定，一方面也顯示「公投」所檢索出的內容歧異性，其中聯合報的判讀參差情況較爲嚴重。爲能有效比對評估，無法認列的文件以及被認列爲「無關」的報導均不作爲文字探勘的比對項，而是以「支持」與「反對」二變項，評估文字探勘的分類效果。

從以上內容分析的結果可知，自由時報的報導意向較爲明確，至少在人工識別上是如此，而聯合報在支持與反對報導的比例就不若自由時報明確，無論這是基於維護平衡報導的形象，或是因爲反對論述較難著力，這種編撰風格上的不同是否造成文字探勘的差異，一般不爲文字探勘評估所重視，但在社科研究上卻應該加以斟酌。

肆、文字探勘的執行

一、關鍵詞選取

依據文字探勘處理程序，首先將六組文本集，透過中央研究院中文斷詞系統 CKIP（Chinese Knowledge Information Processing）斷詞，[4] 並建立「文件—詞彙」矩陣。再利用 TF-IDF 法（Salton and Buckley, 1988: 516-517），計算詞彙權重，抽取出重要的關鍵詞，以減少維度並導入文字探勘。直覺上，詞彙使用次數似乎代表它在文件集的重要程度，但是，若高詞頻的詞彙僅集中在少數文件中，便無法代表整個文件集的意義。TF-IDF 法一方面考量詞頻（term frequency），計算在文件 d_j 中詞彙 t_i 的出現數，即 tf_{ij}，另一方面又考量詞彙 t_i 在文件集中的出現頻率，取其反函數（inverse document frequency），即 $idf_i = \log |D| / |\{d | t_i \in d\}|$，$|D|$ 爲文件總數，$|\{d | t_i \in d\}|$ 爲 t_i 出現在文件集的數目，因此 $TF\text{-}IDF = tf_{ij} \cdot id_{fi}$。本研究將文件集中個別詞彙的 TF-IDF 值加總視爲權重。由於名詞承載意義較爲豐富，經刪除沒意義的停字詞以及單一字，本研究取 TF-IDF 值大於 0.2 的名詞作爲關鍵詞，就六組資料各建立一

4 中央研究院中文斷詞系統 CKIP，http://ckipsvr.iis.sinica.edu.tw/。（中央研究院資訊所，2003）

個「文件—詞彙」矩陣，其中矩陣元素所標記的是關鍵詞在所有報導文件的出現次數，表 2 顯示提取關鍵詞的結果，此時，納入分析的文件數總計 855 筆，關鍵詞數總計 1581 個。

表 2：分析語料的文件數與關鍵詞數

	月份	文件數	關鍵詞數
自由時報	2007.11	106	125
	2008.01	199	336
	2008.03	129	155
聯合報	2007.11	149	229
	2008.01	163	533
	2008.03	109	203

註：月份 2007.11 代表 2007.10.22～11.21，餘依此類推。

二、分類器：支持向量機與簡易貝式分類器

本文採用監督式學習中的支持向量機與簡易貝式分類器執行文件分類任務。支持向量機屬於較新的一種機器學習分類器，可用於非線性分類（Russell, Stuart and Norvig 著，歐崇明等編譯，2011: 18_42-18_46; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000）。其基本概念是，設想文件節點散佈在一個二維平面上，倘若二種文件區隔是線性的，即可在二群節點間尋求一條分隔線，並在這分隔線平行二側再各畫上一條線，使其分別通過二側最接近中央分隔線的若干節點，這些節點就是支持向量。二側平行線間距離達到最大時表示分類區隔最好，此時所找到的最佳分隔稱做最佳分離超平面（optimal separating hyperplane），求解的分類函數爲 $f(x)=sign(\sum_{sv}\alpha_i y_i K(x_i \cdot x)+b)$，其中 x_i 爲支持向量，α_i 是用以解決對偶適化問題，b 爲常數。當二種節點分布交錯，則改以柔性邊界（soft margin）求解，若是分類爲非線性，則映射至高維度特徵空間，尋求最佳分離超平面。由於演算方法有持續改良的

空間，本研究不擬做細部調校。[5] 不過，支持向量機所耗費的訓練時間較長，爲瞭解不同工具或流程對分類效果的影響，本文另使用簡易貝式分類器，做比較測試。

簡易貝式分類器早已應用在各種資料探勘之上，目前仍是文字探勘的基本工具之一，其優點是僅需小量訓練資料，速度快，二分類的效果佳，但其改良模式也不斷被提出討論，本研究作爲社科分析應用，並不著重在優化分類器的討論，而是觀察不同探勘工具的影響。簡易貝式分類器是以貝式定理爲基礎的機率分類器，簡易貝式分類器的「簡易」二字有時又稱爲單純、素樸或天眞，都是指它假設各分類的表徵屬性，其彼此間爲條件獨立（conditional independence），這看似不務實的假設，有時產生不錯的探勘效果。推導上，它透過訓練資料以事後機率取代事前機率後，進行分類預測，可表示爲 $P(C_j|x)=\prod_{i=1}^{n} P(x_i|C_j)\times P(C_i)$（Russell, Stuart and Norvig 著，歐崇明等編譯，2011: 20_7; Tufféry, 2011: 492-497）。其中 $x=(x_1, x_2, \ldots, x_n)$ 代表某特定資料的屬性集合，其中各個屬性各自獨立，C 表示文件類別，因此 $P(C_j|x)$ 屬性分布中出現 C_i 分類的機率，N 爲資料訓練後的特徵屬性總數，$P(x_i|C_j)$ 指不同屬性出現在 C_i 的機率。分類器訓練資料時將屬性值指派給最高機率值的類別，另爲求點簡化，實務上是取對數進行運算。[6]

三、評估方法

本研究以混亂矩陣（confusion matrix）定義文件分類的精確率（Kohavi and Provost, 1998），如表 3 所示，其中 TP 顯示機器與人工對於正例判讀的

5 本文中支持向量機以線性核函數執行學習與泛化，SVM type=C-SVC classification，參數設定爲，cache size=40，cost=1，degree=3，eps=0.001，loss=0.1，nu=0.5，seed=1。屬性選取同樣採支持向量機，其中 attributes to eliminate per iteration=1，complexity parameter=1，eps parameter=1.0E−25，tolerance parameter=1.0E−10，屬性之分類辨識能力經個別評估後加以排序，再以此排序訊息供分類器執行文件分類。

6 本研究在執行簡易貝式分類器時，屬性是先經過選取後再匯入分類器，選取法採最能預測特徵的屬性子集，並去除辨識能力接近的屬性，參數設定爲，thread=1，thread pool=1；另子集空間搜尋是以回溯法強化之 greedy hill climbing 執行，其中 lookup catch size=1，search termination=5。

表 3：混亂矩陣

	機器學習分類正例	機器學習分類負例
人工判讀正例	True positive (TP)	False positive (FP)
人工判讀負例	False negative (FN)	True negative (TN)

資料來源：整理自 Kohavi and Provost, 1998。

一致性個數，TN 表示二者在負例判讀上的一致性個數，FP 與 FN 各表達二者間的差異個數，召回率 $R=TP/(TP+FP)$，精確率 $P=TP/(TP+FN)$，前者代表機器分類成效，後者代表機器分類精確率，若要同時考慮二者，則使用 F 測試值，即 $F=P\times R\times 2/P+R$。本研究以加權平均之精確率與 F 評估值作爲評估與討論依據。

本研究將經過人工判讀的資料，以隨機方式做二區分，一爲訓練資料，一爲受測資料。訓練資料的分類訊息作爲機器訓練用，但不混入受測資料，避免發生過度配適（overfitting）問題（Witten et al., 2011: 32-33）。受測資料的人工判讀結果則作爲比對與評估依據。訓練與受測資料是隨機選取，並採用十折交叉驗證法（10 fold cross validation），亦即隨機切割資料爲十份，再隨機取其中九份爲訓練資料，所餘一份作爲受測資料，隨機重覆十次後，以平均精確率決定最佳分類模型（Kohavi, 1995）。

在實際應用上，研究者可從一個資料集中隨機選取小部份資料進行訓練與評估，並以評估後的模型適用在未標記的大量目標資料。通常用來訓練的資料量愈大，模型評估所得精確率愈高，當然這會導致人力與時間成本增加，而且精確率提升會在某個資料量閾值上發生衰減現象。[7] 因此，本研究在評估分類效果的同時，也將觀察檢索詞、資料編撰、關鍵詞數、文件數、類別大小對於文字探勘的影響。

7 Witten et al., 2011: 149。另 Hopkins 與 King（2010: 241-242）在他們的自動化無母數內容分析模型中，建議在社科研究上，只要取 100 件以上進行人工判讀即可，此時，分類的平均標準誤即可達到 0.04 以下，取 500 件以上對於標準誤的提升沒有太大幫助，不合於成本考量。不過此一模型只用來區別不同類別的百分比，而非用於個別文件的分類。

四、評估結果

利用監督式學習分類器，以訓練資料建立模型，再以此模型就受測資料進行分類，並以十折交叉驗證法，得出文字探勘在文件分類上的精確率，如表 4 所示，當然，此處的精確率意指機器與人工分類的吻合度，不假設其中有「眞」值。依據表 4 評估結果，首先明確顯示無論使用何種分類器，對自由時報文件集的分類效果均優於聯合報。這是否意味著，檢索詞所擷取的二報資料內容差異影響文字探勘效果，有必要做進一步解析。

再以分類器效果來說，支持向量機對自由時報三個月份資料分別達到 86.6%、90.6%、91.4 的精確率，也就說，與內容分析的差距都在 15% 以內，三個精確率數值水準也頗爲一致，以目前常見的文字探勘評估值來看，是相當不錯的表現。在聯合報方面，表現就不那麼理想，距離選舉日的第五、三、一月報導，文件分類精確率分別爲 67.8%、72.6%、70.9%。簡易貝式類器對自由時報文件集分類也有較佳的表現。若與支持向量機相較，除了在聯合報 2007.11 的文件集上有相同表現外，另對二個文件集的分類精確率優於支持向量機，分別爲自由時報 2007.11 的 90.3%，以及聯合報 2008.03 的 82.5%，餘三個文件集均略遜於支持向量機的表現。二種分類器對於六份文件集的分類效果各有所長，整體來說，支持向量機的表現略佳，但差距並不大，差距

表 4：文件分類的效果評估

自由時報			聯合報		
文件集	SVM	Naïve Bayes	文件集	SVM	Naïve Bayes
2007.11	86.6% (85.3%)	90.3% (89.9%)	2007.11	67.8% (68.7%)	67.8% (67.3%)
2008.01	90.6% (91.2%)	87.9% (87.4%)	2008.01	72.6% (72.8%)	64.6% (65.7%)
2008.03	91.4% (91%)	84.3% (84%)	2008.03	70.9% (70%)	82.5% (82.3%)

註：所列數值爲加權平均精確率，括號內數值爲 F 評估值。

較大的文件集，一爲聯合報 2008.01 文件集，此處支持向量機有 8% 的較高精確率，二爲聯合報 2008.03，這裡簡易貝式分類器有 11.6% 的較高表現。

由於，實務上文件分類仍可支援其他知識挖掘之用，因此，除了關注精確率之外，也可將召回率納入考慮，亦即確保精確率能高於 F 值，或至少不要低於 F 值過大，以使分類器對於每個類別的特徵都具備良好的辨識力。表 4 中的 F 值顯示，支持向量機有三個文件集的精確率略低於 F 值，也就說分類器的分類辨識能力並不能完全反應在精確率上；而簡易貝式分類器則有五個文件集的精確率高於 F 值，表現較優。由於表 1 已顯示，自由時報的類別筆數相較於聯合報有較大的懸殊差距，若再比對表 4 的 F 值，意味著支持向量機較易受到類別筆數大小影響，對簡易貝式分類器的影響則不明顯。顯然，不同分類器對於類別筆數的敏感度不同，社科研究不乏遭遇大小類別懸殊情形，反覆測試分類器或增加訓練資料量是比較穩妥的作法。不過此處整體來說，二種分類器的精確率與 F 值都相當接近，表現尚稱穩定。

另一個值得觀察的是，文意分類效果的一致性。從表 4 可知，支持向量機無論對自由時報或聯合報的文件集，其分類效果均相對穩定，也就是說在相同的檢索條件下，相同來源的不同文件區分之間，都有相對穩定的文意分類效果，表示分類模型對於複雜多變文意的掌握能力佳。相較之下，簡易貝式分類器對於聯合報文件集的效果差距較大，也因爲有這樣差距存在，實務上對於訓練與受測資料的選取必需謹慎。最後，比較表 2 與 4，發現二種分類器的精確率，與文件數、關鍵詞數間的關聯頗爲分歧，經以 Pearson 相關加以檢證，發現其間關聯均不顯著（$p>0.05$，表格從略），亦即，在本研究的實例中文件數與關鍵詞數尚不影響文字探勘的效果。

伍、討論與建議

依據前述的評估結果，本節就分類效果做進一步評述，並討論二報分類效果差異的問題，最後從社會科學應用角度提出文字探勘應用的相關建議。

一、文意分類效果評述

由於文字探勘與內容分析所指範疇極爲廣泛，在對前節評估效果做進一步討論前，謹就本研究的設計考量與資料特性再做簡要的強調。

首先，本文是以「文意區別」爲目的所進行之文字探勘驗證，與常見以「主題」爲標的的文件分類仍有不同，困難度相對較高，目前在中文文獻中仍不多見。其次，本文在設計上爲求簡明，以二分類做區分，使分析得以聚焦在文字探勘文意區別的可行性，這雖是探索階段的權宜做法，但在社科研究上仍有實務上的需求。最後，本研究不著墨文字探勘前端的資料處理（如斷詞系統）與系統組合（如語法剖析），或是後端的模型創建或決策支援，但強調從過往單純的內容分析電腦輔助，過渡到運用一般文字探勘程序，也就是希望在一個共通的介面上，以社會文本分析社會議題，聚焦文件分類的技術與效果，以提供領域內外的評估參照與經驗累積。

前已論及，以「公投」檢索出的二報文件集內容極爲複雜，包含多元的意義表達，這主要是因爲，新聞報導是一種自然語言，「公投」一詞所檢索出的報導可能包含許多不同的意義表達與外在事件描述，這不僅涉及書寫者的用詞及所欲表達的意念，也涉及報導立場與取材方式，再加上「公投」檢索已擷取出諸多容易混淆的公投案與相關論述，因此意圖判讀單一意向（支持民進黨或國民黨版公投）自然是不容易的。在第參節執行內容分析時已指出人工判讀參差占總筆數最高達 24.8%（見表 1），表明了文本的複雜度。然而，這些複雜與困難其實也正是社科研究經常遭遇的情境。

在前述的設計考量與複雜文本條件下，本研究對監督式學習分類器的評估顯示，納入訓練與測試的文件數、關鍵詞數、類別筆數對於二種分類器的影響並不大，支持向量機在二報精確率水準都能保持穩定，簡易貝式分類器在聯合報文件集則有較大起伏，若再考量精確率，本實例以支持向量機的表現較佳。這也表明不同分類器對不同資料特性的掌握各有所長，足以影響文意分類的效果，研究者應反覆多方測試，選擇最能掌握手中文件特性的分類器。近期發展成熟的「整合選擇」（Ensemble Selection）流程，已可比較整合不同演算方法的效果（Caruana et al., 2006; Caruana et al., 2004），足以提升文

件分類的效能，而當未標記資料過大時也可以考慮使用「準監督式學習」（semi-supervised learning），以節省人力與時間成本（Witten et al., 2011: 294-296）。

就文件分類來說，當然希望分類器評估結果可以獲得較高的精確率。一般文獻通常視精確率數值愈高愈好，並做爲比較參考的數值。但在相關文獻討論中可以見到 50% 至 90% 的數值結果（李政儒等，2012；曾元顯，2002），差距範圍頗大。惟論者也指出，對於較複雜的主題分類任務，F 值達到 80% 以上可視爲相當高的數據（曾元顯，2002: 76）。以本研究所投入的有限文本規模與複雜文意區別來說，在考慮 F 值的情況下，精確率達到 80% 以上當可視爲是較高數值。據此，本研究中二種分類器對自由時報的精確率均佳，意味著文字探勘對於複雜意向的文件分類是有效的，以機器學習減少人工判讀成本具備一定的理據與效用。

至於對聯合報文件集的較低精確率，除了分類器因素，也與文件內容複雜度有關。若人工判讀已存有較大差異，機器判讀也不可能太高，前文已從認識論角度有所析論，而曾元顯（2002: 76）也指出，分類器與人工分類的不一致性，若達到不同人工分類之間的不一致性，即可視爲最佳的分類狀況。據此，儘管納入測試的二種分類器對於聯合報的分類效果稍差，但若考量本實例人工判讀的參差比例，尚不能就此否定文字探勘在文意判讀的效果。由於支持向量機對聯合報三個文件集的精確率相對穩定（67.8%～72.6 之間），提供了進一步調校的空間。至於簡易貝式分類效果的不一致，主要是因爲對聯合報 2008.03 文件集有較突出的表現，這除了是分類器的性能關係，宜考量以隨機方式篩選建模資料，或者在初期資料處理階段，建立核心關鍵詞庫，以求關鍵詞能更完整反映資料特徵，貼近所欲挖掘的訊息標的。至於資料特性本身所構成的影響將於下文做進一步討論。

必須強調的是，本研究在設計上爲壓迫機器學習的空間，並考慮社科研究可能遭遇的狀況，納入分析的文件集篇數都不多，在如此條件下，監督式機器學習在分類效果與穩定性上都值得肯定。這連帶支持了文字探勘的假設基礎，亦即人類語言的描述確實存在某種結構，儘管這種結構隱而不顯，也參雜著許多變異與雜訊，但只要使用的方法得當，仍可以透過文字探勘取得

不錯的區別效果，當運用在巨量資料時，基於減省時間與人力考量，可以彌補傳統內容分析之不足，或是達成內容分析無法做到的分析工作。由於本研究並未對分類器做細部調校，相信隨著演算方法的進步，會有更多的改善空間。

二、編撰風格及其網絡結構

那麼，如何解釋自由時報與聯合報資料的精確率落差呢？儘管以「公投」進行檢索的用意在收納多元複雜文本，在相同的資料處理與探勘程序下，表 4 中二報在精確率上的差距，意味這種差距也與二報的表述特徵有關。此一發現，使得文字探勘的應用問題，從外顯詞意與內隱意向的區別，深化到文本產製的差異，此一差距可稱做是「編撰風格」的不同。編撰風格的內涵與成因超過本文的析論範疇，但新聞學研究卻有類似的概念，亦即「新聞風格」。

「新聞風格」指「在特定社會文化環境中，新聞媒體在組織文化影響下，展現於組織常規以及語言常規方面的獨特特徵。」（施祖琪、臧國仁，2003: 153）其中組織常規包括新聞價值、讀者定位、路線分配、時間規劃等，語言常規指的是新聞結構、修辭技巧等。新聞風格強調了文本產製問題，亦即媒體「如何」表現其「獨特性」（臧國仁、施祖琪，1999: 11），也就說，對相同事件可以有不同的報導方式，從而影響新聞內涵與詞語使用。若將新聞風格加以延伸，泛指一般的資料產製與文件編撰，則本文所指編撰風格可定義爲資料產製受到組織文化影響，所顯現的組織性詞語特徵。本文所強調的編撰風格並不侷限在新聞報導，因此，是以有別於其他同類型資料，甚或是研究者規範認知，所顯現的組織性詞語特徵。以本研究的資料來說，在「公投」議題上，基於組織立場、用人政策、採編規範、論述策略、素材選擇、訴求對象、書寫手法、議題營造等差異，二報所呈現的文字表述與取材內容，對特定議題產生或多元或收斂的表達差異，進而影響文字探勘的效果。

這種源於編撰風格所構成的影響可以明顯地在二報資料中觀察出來。由表 1 可知，人工判讀認列爲「無關」（含內容中立、無關或無法判斷）所占的比例，自由時報三月平均爲 18.62%，聯合報平均爲 37.57%，顯然聯合報的文字表述更多元而分歧，或說自由時報顯得更一致而收斂。若以帶入文字

探勘的二分類（即支持與反對）總和所占百分比進行比較，如表 1 所示，可以看出二報報導立場上的差別。自由時報對於民進黨版公投呈現極爲鮮明的支持立場，而聯合報儘管如一般預期，報導反對民進黨版公投（亦即支持國民黨版公投）的篇幅較多，但支持與反對比例相較而言緩和許多，如此經由機器學習，聯合報用詞與表達上的多元與發散，自然造成其文件分類效果不若自由時報來得精確。

在編撰風格的影響下，主題選擇與語意表述會以獨特的方式鑲嵌在自然語言結構中，並在文字探勘的處理過程裡，轉化爲向量空間的特徵差異。爲了簡明表達這種差異，表 5 將這種「文件—詞彙」向量空間映射在一個雙模網絡（two-mode network）中，以網絡結構的角度與指標呈現二報的特徵差別。[8] 首先聚焦網絡的密度，依據表 5 數據，顯然自由時報的文件或詞彙間連結程度是相對較高的；其次，平均距離指網絡節點間最短距離的平均值，表 5 的數值表明聯合報的詞彙間連結需要較大的距離；網圖的破碎性表示網絡次級成分的不可及程度，表 5 顯示聯合報的數值均高於自由時報，代表擁有較多的獨立性議題；最後，遞移性網絡指標在呈現網絡的重要基本結構，亦即三元組（triad）的連結程度，表 5 數值顯示自由時報遞移性相對較高，表示其文件或詞彙之間，不僅兩兩連結較緊密，而且與任一文件或詞彙有連結的另二個文件或詞彙間，其關聯性也相對較高。以上分析指出，自由時報的報導與用詞呈現較好的連結，而聯合報顯得較爲鬆散，此一特徵差異存在於文字探勘執行前的「文件—詞彙」結構中，本研究推斷正是這種結構差異影響了文字探勘的文意辨識。

爲驗證上述主張，進一步將「文件—詞彙」網絡轉換爲「詞彙—詞彙」

8「文件—詞彙」是以矩陣形式，作爲向量空間模型（Vector Space Model）的初始資料（Turchi et al., 2009: 11-12）。該矩陣以所蒐集的「文件」筆數爲橫列，以經過選取可表達整體文件集特徵的關鍵詞彙爲直行所構成，將該矩陣映射成網絡圖時，因同時呈現二個維度，故而稱做「雙模網絡」（Hanneman and Riddle 著，陳世榮譯，2013: 284-285）。雙模網絡利用網絡理論即可計算出其基本的網絡特徵與指標。其中，密度是網絡連線除以可能連線的數值，平均距離是節點間測地距離的平均數，破碎性是網絡中不可及的成員比例，遞移性呈現任三個節點形成三元組（形成連結）的機率，以上數值均同時考量雙模的規模（Borgatti and Everett, 1997）。

表 5：二報「文件—詞彙」向量空間之網絡特徵比較

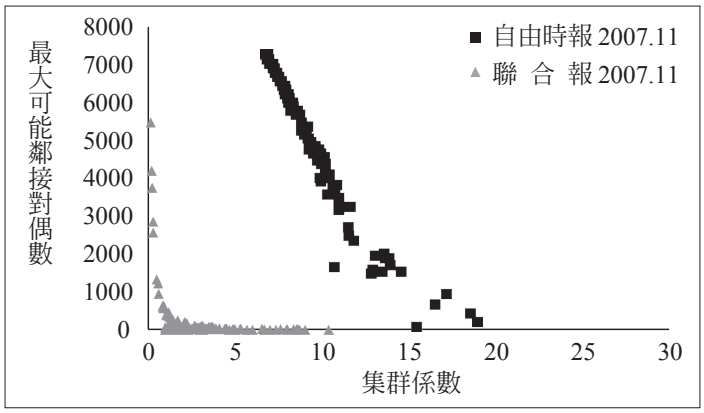
	密度	平均距離	破碎性	遞移性
自由時報 2007.11	0.21	2.40	0.02	0.38
自由時報 2008.01	0.11	2.68	0.01	0.28
自由時報 2008.03	0.17	2.48	0.02	0.38
聯 合 報 2007.11	0.02	3.76	0.34	0.32
聯 合 報 2008.01	0.06	3.11	0.10	0.26
聯 合 報 2008.03	0.13	2.71	0.08	0.35

註：各網絡指標演算方法參考註 8。

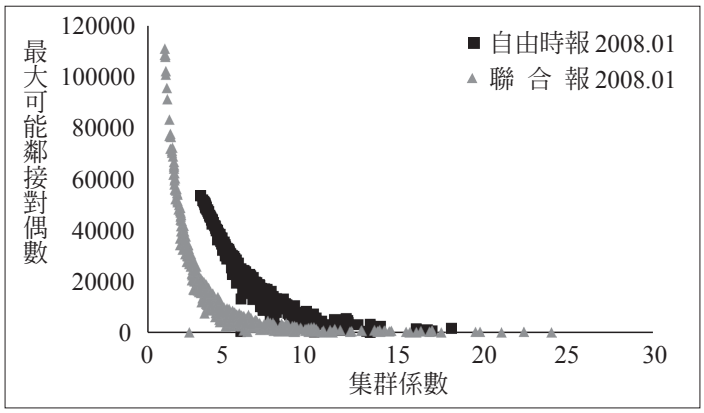
的共詞網絡（co-word network），並比較二報關鍵詞彙的集群係數（cluster coefficient）分布模式。[9] Watts 與 Strogatz 以高集群係數與低平均距離，證明「小世界網絡」的存在，以無向網來說（例如共詞網絡），個體的集群係數可表示爲 $C_n=2e_n/(k_n(k_n-1))$，其中 k_n 代表 n 節點的鄰接數，$k_n(k_n-1)$ 代表所有這些鄰接點間的可能連結總數，e_n 代表 n 點在所有鄰接點間的連結數，因此個體的集群係數是某節點的鄰域（neighbors）連結數，以及與此一鄰域最大可能連結的比例（Watts and Strogatz, 1998）。此處，共詞網絡中的節點是描述文件集特徵的關鍵詞彙，因此詞彙間集群傾向，一方面有連通不同文件的意涵，一方面也代表不同議題或主題的連結，所以，集群係數表達了一個詞彙之跨文件、跨主題的串聯程度。如以圖 1 所示，Y 軸數值代表一個詞彙連結足以構成完整子圖的最大可能對偶數，X 軸表集群係數，當詞彙鄰域最大可能連結極大，但集群係數小，表示該詞彙的意義連結多元而發散，反之，當鄰域最大對偶數少但集群係數大，代表該關鍵詞的意義指涉相當集中且收斂。圖 1 分別將二報同月資料做兩兩比對，可以明顯看出 2007.11 自由時報

9「文件—詞彙」矩陣經轉換爲「詞彙—詞彙」矩陣後，矩陣元素（細格）代表不同詞彙在相同文件中的共現次數，因此「詞彙—詞彙」矩陣在描述一組文件集所含不同關鍵詞間，兩兩共現的訊息（有關雙模轉換請參閱，Hanneman and Riddle 著，陳世榮譯，2013: 98）。當「詞彙—詞彙」矩陣以網絡圖呈現，網中的節點代表關鍵詞彙，而連線表示詞彙的共現連結。以此一共詞網絡爲基礎，即可尋求單一詞彙的集群係數，該係數代表單一詞彙與鄰接詞彙能形成完整子圖（clique，小團體）的程度。

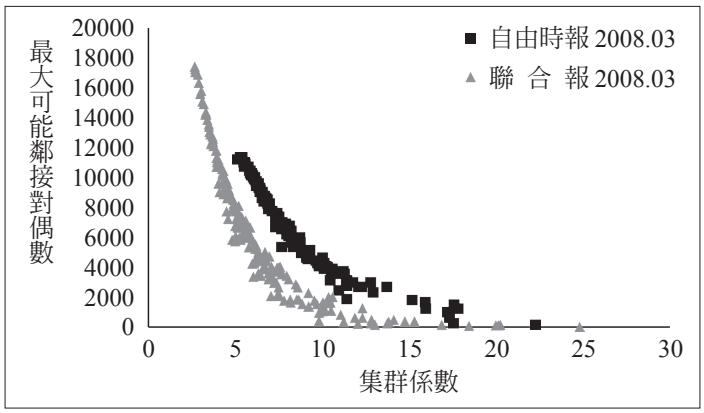
圖 1：二報共詞網絡的個體集群係數比較



1-1



1-2



1-3

資料來源：本研究整理分析。

文件集中所有詞彙的集群係數明顯高於聯合報詞彙（圖 1-1），2008.01 及 2008.03 資料中，聯合報有不少關鍵詞雖與其他詞彙擁有較多的連結機會（Y 軸），但實際構成連結的數目，亦即集群係數，均相對低於自由時報（圖 1-2, 1-3）。換言之，聯合報中有關公投的表述文字，其意義串聯程度相對低於自由時報，這便影響了機器學習與文件分類的效果。統合比較也可以看出，自由時報三個月份報導文字的集群係數主要落在 5 至 15 之間，可見其關鍵詞的意義串聯相當一致而穩定，相對地，聯合報三個月報導集群係數分布變異較大且不一致。值得注意的是，聯合報在選前一個月的報導，相較於之前二個月集群係數略爲升高（右偏），並爲簡易貝式分類器所捕捉（見表 4），顯示詞意串聯程度與文字探勘精確率有一定的關聯。

綜合以上，足以說明相同的檢索條件與探勘程序下，文件分類效果之所以有優劣之分，正是因爲初始資料的編撰風格差異所致，這種差異顯示在詞彙向量空間的網絡結構疏密，以及詞彙意義串聯的高低之差。

三、延伸討論與建議

以下從編撰風格所透顯的資料特性問題出發，就社會科學運用文字探勘進行文意區別所可能面對的問題，做一番延伸性的討論並提出建議。

本文第貳節已述及文本產製因素向來爲內容分析法所重視，而前段的分析也指出文字探勘應用不能忽視這一問題。資料編撰風格的癥結，並不在二報的落差，因爲實際應用上未必有需要或有機會去比較二種文本風格的差異，而是說任何文本都有編撰風格存在，甚至研究者對於特定資料的認知也同樣受到文化與制度的侷限，當文本的特殊表述風格，與理論框架或研究者判斷有差距時，便容易影響文字探勘的效果，尤其需要以檢索方式獲取資料時更容易發生這一問題，因爲檢索策略更容易參雜主觀判斷。解決之道在於，對自己所納入分析的文本，保持質疑精神與彈性，一方面反覆斟酌所擷取資料與研究目的的關聯性，二方面可先利用參考文件建立核心詞庫，並從反覆測試與模型評估中，檢視文字探勘效果與原先預期是否存在差距。例如，本研究的評估結果顯示，「公投」關鍵詞並不能清晰地擷取出聯合報有關支持或反對某黨公投案的報導意向，有必要彈性嘗試其他檢索字詞，搜尋

出能清晰表達意向的文件集。

此外，資料內容與研究目的的連結問題，也必須考量監督式學習的需求。本文已指出不同類別的文件筆數須能滿足分類器學習的需求，否則必須尋求適合挖掘該文件特徵的分類器，在本研究中，簡易貝式分類器較不受大小類別的影響。由此延伸出的另一個問題是，要確保內容篇幅對分類目標的涵蓋程度，如果所欲探查的意向，其相關表述相對於文件篇幅（如長篇演說）相當稀疏，將不利於機器學習；若無法避免，則有必要加大訓練資料的量，對於訓練資料的選擇也必須更加周延（Leetaru, 2012: 78；陳文華等，2003: 53-54）。另有鑑於本研究中，簡易貝式分類器對於聯合報不同月份的分類精確率並不一致，除了應注意資料的適切性外，建議在資料準備階段，以隨機抽樣方式選取建模資料；若文字探勘有明確的訊息需求標的，如政策支援，建議先行建立能反映研究課題（或資料特徵）的參考文件或核心詞庫（Laver and Garry, 2000），以使監督學習發揮最大效能。

資料特性與風格之所以重要，還在於社科研究者經常必須面對開放式社會文本，例如新聞或社群媒體（social media）文字，在這樣的資料來源中，文本常包含多元意涵，主題與意向經常隨著事件演變而變化，對於長時間、大量的文件文意區別構成相當大的挑戰。此所以在中文文獻中，常見文字探勘應用在科技或專業文本的分析（尹其言、楊建民，2010；林頌堅，2010；施百俊、施如齡，2006；戚玉樑、蔡明宏，2007；賴志遠等，2009），因爲專業文本中的語意維度是較爲單純的；即便有針對開放式社會文本進行分析，分類標的也多集中在主題與類型的判別（許中川、陳景揆，2001；陳文華等，2003），而少文意區別。以本文驗證結果來說，結合傳統內容分析與監督式學習，不失爲文意區別分析的可行途徑，尤其當意義萃取被視爲未來文字探勘領域的發展重點時（Miner et al. eds., 2012: 996），內容分析與文字探勘間的互補性應加以重視。

前端處理與系統整合，雖不在本實驗設計的範疇，但對於文字探勘的文意判讀也至關重要。如果說，計算語言技術對社科研究的幫助，主要發揮在前端資料處理的自動選碼與標記，以及後續的自動訊息萃取（Alexa, 1997: 14），則如何透過專家審定，持續提升斷詞系統、通用詞庫與領域本體的涵蓋

面與精確性，將決定文字探勘的語意徵別效能。Franzosi 指出，未來如何將語言學，尤其是敘事理論（theories of narrative）與修辭學，納入文字分析系統中將是關鍵性工作（Franzosi, 2008: xxxii-xxxv）。尤其當前混合訊息（mixed message）與曖昧用語對政治社會持續發揮著既幽微而有力的影響時（Luck, 1999; Rockwell, 2006），如何解析這樣的溝通與意向，仍需仰賴跨領域專家的合作，社會科學界自然也不應缺席。除了前端資料處理，操作流程階段的系統整合也很重要，例如融入情緒分析（李政儒等，2012）、文字蘊涵（楊善順等，2013）、或語法剖析（如 The Stanford Parser）等系統，也是提升機器文意判讀的重要輔助工具。

不過，當系統與演算技術愈爲精進，專業區隔與知識門檻也愈高，愈容易妨礙技術的應用與擴散。Franzosi 就曾指出電腦輔助儘管帶來甚多益處，但未必能減少時間成本（Franzosi, 2008: xxxv）。尤其在個別的、特定議題的、或任務導向的文字探勘上，共通系統並不一定能滿足特殊需求，研究者往往需要掌握新字詞、長字詞、左右字串（林頌堅，2010；許中川、陳景揆，2001），建立或增補辭庫（Laver and Garry, 2000），或做必要的系統整合（Junqué de Fortuny et al., 2012），這些特定、進階的需求，對社科研究者來說均構成額外的技術門檻與時間耗費。原則上，文字探勘系統應滿足精確、可靠以及有用的原則（Miner et al. eds., 2012: 994），以目前中文繁體相關系統來說，除了滿足上述原則，另應致力於透明開放，強化友善使用的程度，以利應用面的擴大。此外，檢視國內文字探勘文獻可知，目前研究者多爲資訊專業領域，社科研究者投入較少，社科學界有必要重視文字探勘技術發展走向以及相關應用的可能性，同時積極人才培育。唯有擴大參與層面，學門領域的需求才能適時地納入技術改良進程，構成良性的反饋與循環。

陸、結論

誠如 Franzosi（2008: xxxv）所言，當傳統的主題與敘事分析已被整合在單一的計量框架進行分析時，文字探勘技術的應用及其相關問題已不容忽視。但是，社會科學研究基於其長久的分析傳統，對於文字探勘應用並不是

沒有疑慮。本文從內容分析與文字探勘的比較切入，試圖釐清問題癥結。首先，有必要認知，文字探勘技術仍處於發展階段，含有實驗科學的性質（Witten et al., 2011: 403），社科學界不需過度苛求，但也不能漠視它的效果與影響。本文進一步從方法程序的比較中發現，二者在方法學上存有共通性，但在認識論上社會科學更強調解讀的彈性。作者因此主張，社會科學仍應善用文字探勘，發揮其優勢，並藉由人工判讀的融入，縮限特定文意定向，保留文件多元意義的可能性。於是，留待考察的問題便著落在機器與人工判讀的差距。本文即在拋磚引玉，建立文字探勘社科應用的經驗依據，從中發現優劣，尋求改善，提供計文字探勘應用的不同考量面向。

本研究以社科研究經常存在的文意區別需求，依循一般文字探勘程序，進行分類器效能的驗證，這正是嘗試在一個共同的介面上，提供不同領域可資參酌的測試經驗，它不致力個別模型的建立，也沒有借助語法剖析或其他輔助系統，而是以詞語特徵爲基礎，透過監督式學習分類器就特定議題進行文意解析。當然，這不意味這項嘗試提供了最佳的文意分類方法，更不表示此一分析程序即能滿足社會科研究上的許多實務需求，諸如政策支援或輿情分析。但本研究的貢獻在於確證，監督式學習分類器足以勝任社會文本的文意區別，而且文本複雜度、文件數、關鍵詞數，與大小類別，對文件分類影響不大，效果值得肯定，這也意味著，在內容分析與文字探勘的互補爲用下，社科研究可善用各種不同電子典藏文本，諸如新聞、記錄、社群媒體等巨量資料，就各種不同社會議題或爭議，進行分類分析與意向調查，以獲取此前力有未逮的知識探索機會。尤其是對於特定議題，社會、政治、政策研究者或機關部門都可以在共同的探勘介面上進行分析，並相互參照、比較、偵錯，一旦應用面擴增，必將帶動新模型的建立，與探勘流程的精進。

當然，由於不同分類器會產生不同的效果，研究者宜評估不同方法以取得最佳的分析模型。此外，本研究也發現文本資料存在編撰風格的問題，資料表述特性可能與研究者的原始設想有差距，進而影響文字探勘的效果，研究者應於資料處理階段，及早發現文件編撰風格對研究目的所構成的影響。爲求聚焦，本研究以文意的二元區別進行分析，但實務上，尚有多元名義或次序尺度的意義分類需求，而所謂「中立」意向又代表何種意義等，都有待

經驗分析做進一步釐清。此外，社會科學中，個別型、獨立性議題探索的需求較大，因此如何透過核心詞庫提升文件特徵的萃取，如何結合情緒或語法剖析滿足不同的文本解讀需求，是另一個值得探索的面向。

附錄：內容分析編碼簿

分析任務：就各單篇新聞報導，研判是否「支持民進黨版或國民黨版公投」，標記執行人員於報導內容合於研判基準之處，標記「支持」、「反對」、「無關」等三種編碼。

標記對象：以「公投」一詞檢索聯合報與自由報電子資料庫中，2008 年總統大選前第一（2008.02.22～2008.03.21）、三（2007.12.22～2008.01.21）、五月（2007.10.22～2007.11.21）之新聞報導。

分析題目：研判報導內容是否「支持民進黨版或國民黨版公投」，但爲使標記意義明確，以民進黨版公投爲基準，亦即，「支持」民進黨版公投等效於反對國民黨版公投，「反對」民進黨版公投等效於支持國民黨版公投，無法判讀、中立、無關公投者列爲「無關」。

編碼選項：支持：支持民進黨版公投案。

反　　對：反對民進黨版公投案。

無　　關：中立、無法判斷、或無關公投。

研判基準：

編碼	標記研判基準
支持（正）	a. 篇章表達支持民進黨版公投案之主觀意見（含執筆人與被報導人） b. 篇章報導支持民進黨版公投案之相關外在事件 c. 篇章報導有關反對民進黨版公投案之相關外在事件，卻意圖表達支持之主觀立場
反對（反）	a. 篇章表達反對民進黨版公投案之主觀意見（含執筆人與被報導人） b. 篇章報導反對民進黨版公投案之相關外在事件 c. 篇章報導有關支持民進黨版公投案之相關外在事件，卻意圖表達反對之主觀立場
無關（無）	a. 內容中立 b. 內容無法判斷支持或反對立場 c. 內容無關公投案
附註	不同階段的相關爭議或事件也代表著正反意見，應一併列入考量。例如，此期間計有第三、四、五、六公投案被提出；一階段或二階段領票也代表對不同公投案的態度；呼籲選民放棄領取公投票等均是。

參考資料

A. 中文部分

Hanneman, Robert A. and Mark Riddle（著），陳世榮（譯）

2013 《社會網絡分析方法：UCINET 的應用》。高雄：巨流。(Hanneman, Robert A. and Mark Riddle, 2013, *Introducition to Social Network Method*. Roger S. Chen (trans.). Kaohsiung: Chuliu.)

Russell, Stuart and Peter Norvig（著），歐崇明、時文中、陳龍（編譯）

2011 《人工智慧：現代方法》，第三版。新北市：全華圖書。(Russell, Stuart and Peter Norvig, 2011, *Artificial Intelligence: A Modern Approach.* (3rd ed.) Chung-ming Ou, Wen-chung Shih, and Long Chen (trans.). New Taipei: OpenTech.)

中央研究院資訊所

2003 《中文斷詞系統》。2013 年 5 月 1 日—2013 年 10 月 31 日，取自 http://ckipsvr.iis.sinica.edu.tw/ (Academia Sinica Institute of Information Science, 2003, *Chinese Knowledge and Information Processing*. Retrieved May 1, 2013–October 31, 2013, from http://ckipsvr.iis.sinica.edu.tw/)

尹其言、楊建民

2010 〈應用文件分群與文字探勘技術於機器學習領域趨勢分析以 SSCI 資料庫爲例〉，《長榮大學學報》14(2): 1–16。(Yin, Chi-yen and Jiann-min Yang, 2010, "Trend Analysis in Machine Learning Research from SSCI Database by Document Clustering Manipulation and Text Mining Methodology," *Journal of Chang Jung Christian University* 14(2): 1–16.)

自由時報

2007–2008 《自由時報電子報》。2013 年 3 月 1 日—2013 年 8 月 31 日，取自 http://news.1tn.com.tw/search (Liberty Times, 2007–2008, *Liberty Times Net*. Retrieved March 1, 2013–August 31, 2013, from http://news.1tn.com.tw/search)

李政儒、游基鑫、陳信希

2012 〈廣義知網詞彙意見極性的預測〉，《中文計算語言學期刊》17(2): 21–36。(Li, Cheng-ru, Chi-hsin Yu, and Hsin-hsi Chen, 2012, "Predicting the Semantic Orientation of Terms in E-HowNet," *Computational Linguistics and Chinese Language Processing* 17(2): 21–36.)

林琬眞、郭宗廷、張桐嘉、顏厥安、陳昭如、林守德

2012 〈利用機器學習於中文法律文件之標記、案件分類及量刑預測〉，《中文計算語言學期刊》17(4): 49–68。(Lin, Wan-chen, Tsung-ting Kuo, Tung-jia Chang, Chueh-an Yen, Chao-ju Chen, and Shou-de Lin, 2012, "Exploiting Machine Learning Models for Chinese Legal Documents Labeling, Case Classification, and Sentencing Prediction," *Computational Linguistics and Chinese Language Processing* 17(4): 49–68.)

林頌堅

2010 〈利用自組織映射圖技術的研究主題視覺呈現及其在資訊傳播學領域的應用〉，《圖書資訊學研究》5(1): 23–49。(Lin, Sung-chien, 2010, "Visual Presentation of Research

Topics with a Self-Organizing Map and Its Application to the Field of Information and Communication," *Journal of Information Communication* 5(1): 23-49.)

俞士汶

2003 《計算語言學概論》。北京：商務印書館。(Yu, Shiwen, 2003, *Introduction to Computational Linguistics*. Beijing: The Commercial Press.)

施百俊、施如齡

2006 〈以文字探勘技術探究部落格之網路媒體特性〉，《淡江人文社會學刊》28: 95-122。(Shih, Bai-jiun and Ju-ling Shih, 2006, "Text Mining Techniques to Explore the Web Media Characteristics of Blogs," *Tamkang Journal of Humanities and Social Sciences* 28: 95-122.)

施祖琪、臧國仁

2003 〈再論風格與新聞風格——以「綜合月刊」爲例〉，《新聞學研究》77: 143-185。(Shih,Tian and Kuo-jen Tsang, 2003, "Reexamining the Concepts of Style and News Style—Analysis of the Scooper Monthly (1968-1982)," *Mass Communication Research* 77: 143-185.)

戚玉樑、蔡明宏

2007 〈以文件爲對象的概念萃取程序建立知識本體的雛型架構〉，《資訊管理學報》14(3): 47-66。(Chi, Yu-liang and Ming-hung Tsai, 2007, "Knowledge Acquisition Approaches for Building Ontological Conceptual Prototypes in Document," *Journal of Information Management* 14(3): 47-66.)

許中川、陳景揆

2001 〈探勘中文新聞文件〉，《資訊管理學報》7(2): 103-122。(Hsu, Chung-chian and Jing-kuei Chen, 2001, "Data Mining in Chinese News Articles," *Journal of Information Management* 7(2): 103-122.)

陳文華、徐聖訓、施人英、吳壽山

2003 〈應用主題地圖於知識整理〉，《圖書資訊學刊》1(1): 37-58。(Chen, Wun-hwa, Shenghsun Hsu, Jen-ying Shih, and Soushan Wu, 2003, "Application of Topic Map on Knowledge Organization," *Journal of Library and Information Studies* 1(1): 37-58.)

曾元顯

2002 〈文件主題自動分類成效因素探討〉，《中國圖書館學會會報》68: 62-83。(Tseng, Yuen-hsien, 2002, "Effectiveness Issues in Automatic Text Categorization," *Library Associaiton of the Republic of China (Taiwan) Reports* 68: 62-83.)

游美惠

2000 〈內容分析、文本分析與論述分析在社會研究的運用〉，《調查研究—方法與應用》8: 5-42。(Yu, Mei-hui, 2000, "Content Analysis, Textual Analysis and Discourse Analysis in Social Research," *Survey Research: Method and Application* 8: 5-42.)

黃居仁、張如瑩、蔡柏生

2004 〈語意網時代的網路華語教學——兼介中英雙語知識本體與領域檢索介面〉，見羅鳳珠（編），《語言，文學與資訊》，頁 461-485。新竹：清華大學出版社。(Huang, Chu-ren, Ru-yng Chang, and Dylan B. S. Tsai, 2004, "Chinese Language Education and the Developing Semantic Web: An Introduction to Chinese—English Bilingual Ontology Interface," pp. 461-485 in Feng-ju Lo (ed.), *Language, Lerature, and Information*. Hsin-

chu: National Tsing Hua University Press.)

楊善順、吳世弘、陳良圃、邱宏昇、楊仁達

2013 〈蘊涵句型分析於改進中文文字蘊涵識別系統〉，《中文計算語言學期刊》18(4): 1-16。(Yang, Shan-shun, Shih-hung Wu, Liang-pu Chen, Hung-sheng Chiu, and Ren-dar Yang, 2013, "Entailment Analysis for Improving Chinese Recognizing Textual Entailment System," *Computational Linguistics and Chinese Language Processing* 18(4): 1-16.)

臧國仁、施祖琪

1999 〈新聞編採手冊與媒介組織特色——風格與新聞風格〉，《新聞學研究》60: 1-38。(Tsang, Kuo-jen and Tina Shih, 1999, "Style and News Style: An Exploration of News Stylebooks," *Mass Communication Research* 60: 1-38.)

賴志遠、王玳琪、吳騏、張嘉珍、葉乃菁

2009 《文字探勘在科技政策研究之應用》。臺北：財團法人國家實驗研究院科技政策研究與資訊中心。(Lie, Chee-yuen, Tai-chi Wang, Chi Wu, Chia-jhen Chang, and Nqi-ching Yeh, 2009, *The Application of Text Mining for Science and Technology Policy Research*. Taipei: Science & Technology Policy Research and Informaiton Center, National Applied Research Laboratories.)

聯合報

2007-2008 《聯合知識庫》。2013 年 3 月 1 日—2013 年 8 月 31 日，取自 http://udndata.com/udn (United Daily News, 2007-2008, *Udndata.com*. Retrieved March 1, 2013-August 31, 2013, from http://udndata.com/udn)

瞿海源

1982 〈論社會科學研究方法的相容性與互補性〉，見瞿海源、蕭新煌（主編），《社會學理論與方法研討會論文集》，頁 245-266。臺北：中央研究院民族學研究所。(Chiu, Hei-yuan, 1982, "On the Compatibility and Complementary of Social Science Research Methods," pp. 245-266 in Hei-yuan Chiu and Hsin-huang Michael Hsiao (eds.), *Essays on Sociological Theories and Methods*. Taipei: Institute of Ethnology, Academia Sinice.)

蘇中信

2012 〈以紮根理論探討台灣商管期刊中內容分析法的類型〉，《人文社會科學研究》6(2): 1-23。(Su, Chung, 2012, "A Typology of Content Analysis for Business and Management Academic Periodical in Taiwan by Grounded Theory," *NPUST Humanities and Social Science Research* 6(2): 1-23.)

B. 外文部分

Alexa, Melina

1997 "Computer-assisted Text Analysis Methodology in the Social Sciences," *ZUMA Arbeitsbericht* No. 97/07: 1-40.

Blake, Catherine

2011 "Text Mining," pp. 123-155 in Blaise Cronin (ed.), *Annual Review of Information Science and Technology, Vol. 45*. Medford, NJ: Information Today.

Bock, Mary A.

2009 "Impressionistic Context Analysis: Word Counting in Popular Media," pp. 38-42 in Klaus H. Krippendorff and Mary A. Bock (eds.), *The Content Analysis Reader*.

Thoundand Oaks, CA: SAGE.

Borgatti, Stephen P. and Matin G. Everett

1997 "Network Analysis of 2-Mode Data," *Social Networks* 19(3): 243–269.

Caruana, Rich, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes

2004 "Ensemble Selection from Libraries of Models," pp. 137–144 in C. E. Brodley (ed.), *Proceedings of the Twenty-first International Conference on Machine Learning* (also see collected version). New York: ACM Press.

Caruana, Rich, Art Munson, and Alexandru Niculescu-Mizil

2006 "Getting the Most Out of Ensemble Selection," pp. 828–833 in ICDM (ed.), *ICDM '06: Proceedings of the Sixth International Conference of Data Mining*. Washington, DC: IEEE Computer Society.

Cortes, Corinna and Vladimir Vapnik

1995 "Support-vector Networks," *Machine Learning* 20(3): 273–297.

Cristianini, Nello and John Shawe-Taylor

2000 *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York: Cambridge University Press.

Feldman, Ronen and James Sanger

2007 *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Franzosi, Roberto

2008 "Content Analysis: Objective, Systematic, and Quantitative Description of Content," pp. xxi-xlx in Roberto Franzosi (ed.), *Content Analysis, Vol. 1.* London: SAGE.

Glaser, Barney G. and Anselm L. Strauss

1967 *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Pub. Co.

Hand, David J.

2006 "Classifier Technology and the Illusion of Progress," *Statistical Science* 21(1): 1–15.

Holsti, Ole R.

1969 *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Hopkins, Daniel J. and Gary King

2010 "A Method of Automated Nonparametric Content Analysis for Social Science," *American Journal of Political Science* 54(1): 229–247.

Junqué de Fortuny, Enric, Tom De Smedt, David Martens, and Walter Daelemans

2012 "Media Coverage in Times of Political Crisis: A Text Mining Approach," *Expert Systems with Applications* 39(14): 11616–11622.

Kohavi, Ron

1995 "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," pp. 1137–1143 in C. S. Mellish (ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2.* San Francisco, CA: Morgan Kaufmann.

Kohavi, Ron and Foster Provost
1998 "Glossary of Terms," *Machine Learning* 30(2–3): 271–274.
Krippendorff, Klaus
2013 *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.
Lasswell, Harold D.
1965 "Why Be Quantitative?" pp. 40–52 in Harold D. Lasswell, Nathan Leites, and Associates (eds.), *Language of Politics: Studies in Quantitative Semantics*. Cambridge, MA: The MIT Press.
Laver, Michael and John Garry
2000 "Estimating Policy Positions from Political Texts," *American Journal of Political Science* 44(3): 619–634.
Leetaru, Kalev Hannes
2012 *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content*. New York: Routledge.
Luck, Edward C.
1999 *Mixed Messages: American Politics and International Organization, 1919–1999*. Washington, DC: Brookings Institution Press.
Miner, Gary, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, and Robert A. Nisbet (eds.)
2012 *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Waltham, MA: Elsevier/Academic Press.
Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan
2002 "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Languate Processing* 10: 79–86.
Rockwell, Patricia A.
2006 *Sarcasm and Other Mixed Messages: The Ambiguous Ways People Use Language*. Lewiston, NY: Edwin Mellen Press.
Salton, Gerard and Christopher Buckley
1988 "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management* 24(5): 513–523.
Sullivan, Dan
2001 *Document Warehousing and Text Mining Techniques for Improving Business Operations, Marketing, and Sales*. New York: John Wiley & Sons.
Tufféry, Stéphane
2011 *Data Mining and Statistics for Dicision Making*. Chichester, UK: John Wiley & Sons.
Turchi, Marco, Alessia Mammone, and Nello Cristianini
2009 "Analysis of Text Patterns Using Kernel Methods," pp. 1–25 in Ashok N. Srivastava and Mehran Sahami (eds.), *Text Mining: Classification, Clustering, and Application*. Boca Raton, FL: CRC Press.
Watts, Duncan J. and Steven Strogatz
1998 "Collective Dynamics of 'Small-World' Networks," *Nature Australia* 393(6684):

440–442.

Witten, Ian H., Eibe Frank, and Mark A. Hall

2011 *Data Mining: Practical Machine Learning Tools and Techniques*. (3rd ed.) Burlington, MA: Morgan Kaufmann.

Yang, Yiming and Xin Liu

1999 "A Re-examination of Text Categorization Methods," pp. 42–49 in F. Gey (ed.), *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*. New York: ACM Press.

Text Mining for Social Studies: Meaning-based Document Classification and Its Problems

Roger S. Chen
Associate Professor
Department of Public Administration and Management,
Chinese Culture University

ABSTRACT

Along with the growing development of electronic information storage, text mining has increasingly gained attention from scholars and practitioners across various disciplines. In response to the need for meaning differentiation in social studies, the study aims to evaluate supervised machine learning classifiers in terms of the performance of document classification. Setting out from the comparison between traditional content analysis and text mining, the evaluation follows a normal procedure of text mining and applies Support Vector Machine and Naïve Bayes classifiers on non-structural, complex social texts extracted from news media. The outcomes of the analysis validate that text mining manages classification well for documents with complex meaning. However, a further co-word network analysis in the study finds that the editing style of data may affect classifiers' performance. It is suggested that, in the early stage of data processing, greater care must be given to the fit between research problems, editing styles, and classifiers.

Key Words: text mining, meaning differentiation, document classification, machine learning, co-word network analysis