部落格文章情感分析之研究

蕭瑞祥
私立淡江大學資訊管理學系
rsshaw@mail.tku.edu.tw

姜青山
中國科學研究院深圳先進技術研究院
qs.jiang@siat.ac.cn

曹金豐
育學資訊
gino@t-seed.com

簡之文
私立淡江大學資訊管理學系
fresh.chien@mail.im.tku.edu.tw

摘要

如何在龐大的網路社群文章中，有效且快速的擷取所要的情感評論，是情感分析(Sentiment Analysis)重要的基礎工作，本研究選擇在語句層級探討，分析文章中的主觀情緒評論語句。嘗試找出主觀情緒語句與非主觀情緒語句的判斷模式。本研究採用系統發展研究法，使用 SVM(Support Vector Machine)工具將主觀情緒語句進行訓練與測試，實驗中再將 SVM 產出的分類與實際分類做比對，計算其準確率做為系統驗證的依據。

根據本研究的實驗發現，依據 SVM 工具產出的分類，若將屬性詞與意見片語之距離容許誤差值設為 3，可大幅提升距離分類準確率，且可減少距離比對的運算成本，而屬性詞類的分類特徵較能區分主觀情緒語句與意見片語距離的關係。

關鍵詞：情感分析、SVM、知網、主觀情緒

壹、 緒論

根據美國市調公司尼爾森發表的「2009 年全球網路消費者調查報告」顯示，有 7 成的消費者相信網友在網路上發表的意見與評價，陌生網友在網路上的意見，影響力有時更勝於商品推出的品牌形象廣告，若某一網友提出負面評價，公司需付出相對更多的成本彌補其造成的品牌形象傷害，因此對行銷公司而言，瞭解網友在網路上發表的商品評論是重要的 (Liu, 2010)。

Ladh Ounis 等人 (2007) 指出，網路部落格能作為作者本身意見、理念和情感的自我展現，能抒發與分享個人的情緒或評論。在這個網路資訊流通快速的時代，網友已廣泛的透過部落格或參與其他社交媒體交換情感評論或意見，發表對事物的想法或心得。根據 Ku(2007)的研究指出，以部落格與新聞的內容比較，通常部落格文章更帶有主觀性評論的內容。但現在部落格有一半以上的文章是來自轉錄文章，並非作者本身撰寫，或是來自商業性廣告的廣告文章。因此，對於有購買商品需求的消費者而言，在搜尋部落格文章時，常需費大量的心力與時間對文章內容進行閱讀與整理，才能找到其所期望閱讀的商品評論資訊。

楊昌樺 (2006)曾經使用以部落格文章中，常用的表情符號為基礎，判別每篇部落格文章的正反面情緒比例，其結果顯示有近 6 成的準確率。在 Wan (2008)的研究中，曾經將中文文章自動翻譯成英文文章，再將中文與英文的情緒分析結果結合在一起計算準確率，其實驗結果提升了中文的情緒分析準確率。

本研究的目的，旨在分析中文部落格文章中出現的主觀評論語句，希望能準確地偵測部落格文章作者的主觀評論語句，同時歸納出一套適用於部落格主觀評論文章的判斷模式，區別出客觀描述與主觀評論的語句，藉此過濾於部落格中出現的廣告、新聞、轉錄等文章，為讀者找出具部落格作者主觀評論性質的文章。

本研究以中文部落格文章為研究對象，並使用中央研究院 CKIP 中文斷詞系統[1]做文章斷詞。情感判斷依據則是使用知網[2]發布的「中文情感分析用詞語集」，且參考相關文獻 (王正豪、李啟菁, 2010)歸納的否定詞庫，分別用以分析文章中的情感指向和程度分級，否定詞性。本研究範圍僅在探討主觀評論語句的偵測正確性和效率，不考慮文章整體情感意向為何。

貳、 文獻探討

本研究相關研究領域為斷詞系統、SVM、情緒分析等，有相關文獻說明如下列各小節。

一、 斷詞系統

[1] http://ckipsvr.iis.sinica.edu.tw/

[2]《知網》是一個以漢語和英語的詞語所代表的概念為描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係為基本內容的常識知識庫。為董振東先生研究逾十載的研究成果，共收錄了超過 50220 個漢語詞語，所涵蓋的概念總量達 62174 個。

詞是最小有意義且可以自由使用的語言單位。中文文章中，不存在英文字詞之間有空格可做為詞與詞之間的分隔依據，因此必須透過斷詞處理以分隔出中文字詞，以做進一步的分析。由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以盡列所有的中文字詞，且根據中央研究院資訊科學所詞庫小組的研究指出，在統計上一篇文章當中約有 3%~5%的詞彙是未知詞，在新聞類型的文章更是明顯。某些類型的未知詞的詞構非常複雜，也不具備強烈的統計特性，因此未知詞的擷取一直是中文語言處理上一個重要且困難的研究難題。中文斷詞在中文的自然語言處理上，是個相當基礎且非常重要的前置處理工作。目前較為普遍的斷詞系統有兩種，一是由中央科學研究院所研發的 CKIP 中文斷詞系統，與中國科學院所研發的 ICTCLAS[3]。其中 CKIP 是用來處理繁體中文 ICTCLAS 則是繁體與簡體中文都可處理 (黃文奇等, 2011)。

本研究使用中央研究院的中文斷詞系統做斷詞處理，此系統為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文分詞系統。包含了約十萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此詞彙庫及定量詞、重疊詞等構詞規則及線上辨識的新詞，並解決分詞歧異問題，含有詞類標記，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類，將每個詞視為文章中的最小單位。

二、 SVM(Support Vector Machine)

支持向量機(SVM)是一個目前被廣泛運用在分類問題上的數學工具,是根據 Vapnik 的 Max Margin Strategy 發展出來的分類器 (1995)。相較於其他傳統分類器，如：決策樹學習(Decision Tree Learning)、最大熵法(Maximum Entropy)等。SVM 有以下明顯的優點：

（一） 即使在高維特徵向量空間下還是能產生好的效能。

（二） 核心函數能將資料映射到更高維的空間而沒有增加計算複雜度。

支持向量機主要的想法是製造一個最佳的平面可以讓訓練範例向量分成兩個類別：正面與反面（Positive and Negative）並且把這個平面的邊界最大化。圖 1 中，黑色實線就是兩個可將資料分成兩類的平面，兩條虛線中間的距離就是邊界（Margin），也就是 SVM 演算法試著最大化的目標。在虛線兩邊的點稱為支持向量（Support Vectors），而且只有在訓練集中的支持向量會影響整個模型的結果。本研究選用台灣大學林智仁教授所開發的 LIBSVM(2012)工具進行訓練分類。

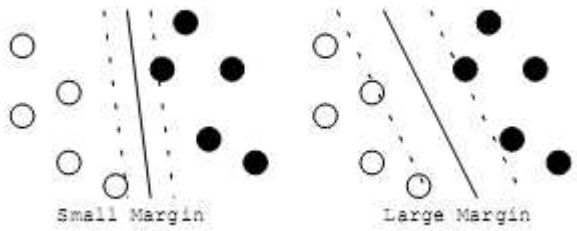

圖 1 兩種可能將資料分開的超平面

資料來源： (Takukudo、Yujimatsumoto, 2000)

[3] http://ictclas.org/

楊等人曾收集部落格中帶有表情符號的句子，訓練 SVM 在語句層級的情緒分類器(楊昌樺等, 2006)。孫等人也曾蒐集在噗浪[4]中帶有表情符號的句子，將情緒區分為正反兩面，使用 SVM 訓練出分類情緒正反兩面的情緒分類器 (孫瑛澤等, 2010)。林的研究使用 SVM 工具，應用在名詞組辨識上，其將判別好的詞分類當成特徵集讓 SVM 分類器參考，辨識出語句的名詞組 (林晏僖等, 2008)。

三、 情緒分析(Sentiment Analysis)

透過文件分類技術可以預測作者發表文章的情緒，常見從文字中偵測情緒問題的方法，找出文字內容和經由人工標記的情緒類別之間的關聯性，當我們所蒐集的語料夠多，使得找出文字和情緒類別之間的關聯具有顯著相關時，就可以利用這些關聯性，預測未知情緒類別的文章之可能情緒 (孫瑛澤, 2010)。關於情緒分析的相關研究，Ku(2007)的研究指出，情感挖掘可區分為三種層級來討論，分別為字詞、語句、文章。大多數的相關研究均在文章層級探討。在楊昌樺(2006)的相關研究中，是根據文章中的表情符號代表作者情感的表達，並將表情符號做了數項情緒分類，如喜、怒、哀、樂，最後指出若將情緒區分為正反兩面來做 SVM 的上的情緒分類器之效率最高。而也有些情緒分析的方法是落在語句層級，既使每一句語句在文章中均表達不同的情緒，但整篇文章在語意上和語法上應會有一致的傾向，如傾向正面情緒或反面情緒。因語句層級的情緒分析是文章全文的基礎，故本研究選擇在語句層級上做探討。楊等人 (楊昌樺等, 2007)曾在語句層級上，選出具有情緒詞彙的語句，透過 SVM 判讀語句的情緒頃向，應用在情緒趨勢的分析上。

有關情感分析有些研究的做法是將文字之間的距離視為判斷依據，如王正豪 (2010)曾在其研究中，以評價詞之文字位置為基礎，計算距離該評價詞 2 個單位距離的程度詞與 4 個單位距離的否定詞，若在範圍之內的即作為對該評價詞的修飾。一般情感分析是以關鍵字與情感字詞之距離，如 Ku (2007)則判定關鍵字與否定詞之單位距離 1 之內的，即視為對關鍵字的評論與修飾行為。Bing Liu (2004)的研究也曾經定義若功能片語與形容詞之字詞距離不超過 3，則將功能片語視為與形容詞是緊密相關的功能片語。不過 Bing Liu(2006)的研究也指出，另一判別情緒之方法為判別比較性質的關鍵字詞，但指出此方法可能將客觀描述性質的比較詞語判別錯誤。在辨別主觀情緒的相關研究中，Pang (2004)曾經利用基於 Minimum Cuts 演算法，透過 SVM 辨識主觀情緒語句。在情感分析另一個重點是判斷字詞，有關字詞判斷以 Bing Liu 研究有兩種方法，本研究是以詞庫方式，並採用大陸「知網」所發展的中文情感分析用詞語集，包含約 9193 個字詞，其內容包括了六個子類別詞語集，各子類別列表出相關字詞如：

- 正面情感字詞（如：贊賞、喝彩…）
- 負面情感字詞（如：後悔、絕望…）
- 正面評價字詞（如：便宜、聰明…）
- 負面評價字詞（如：笨重、低劣…）
- 程度級別字詞（如：超級|過度、最級|非常、很級|格外、較級|更為、稍級|蠻、欠級|半點…）

[4] http://www.plurk.com

l　主張字詞（如：感知l發覺、認為l相信...）

主張字詞目的在於表達人有所感知，見解而使用的字詞，例如「我『發覺』功能其實很不錯」，與「我『相信』它是很耐看的」，然而，忽略這些字詞並不影響分析意見的成效，依然可以從「功能其實很不錯」與「它是很耐看的」取得意見資訊。再者，現代許多人的書寫習慣，在描述自己內心的想法時，有時反而忽略這類字詞，，例如「他好可愛喔」，而不說「我覺得他好可愛喔」 (王正豪、李啟菁, 2010)，故在本研究中，主張字詞將不引用進行分析。而高 (2007)的研究是以詞彙語意學的觀點，整合了包含知網在內的詞彙知識庫，找出任兩個詞彙間的關係，如同義、反義、上下位關係等等。

Bing Liu (2008)在其研究中定義了評論模型，每個被評論的物件可能是產品、服務、主題、個人、組織或事件，並有其所屬的一組被評價的屬性，物件可以再被視為另一物件的子物件。而 Hu (2004)曾經在其研究中，選擇在語句層級上分析出對於產品屬性的正面評論與反面評論，再將所有分析出的評論做總結。Bing Liu 等人(2005)曾經提出一個視覺化檢視情感分析結果的意見觀察雛型系統，能有效幫助使用者或組織看到產品的優點與缺點。

參、　研究方法

本研究採用 Nunamaker (1991)等人所提於資管領域之研究方法中的系統發展研究法。系統發展研究法包含了 5 個重要的研究流程（如圖 2 所示），依序是：

一、　建構概念框架：包含陳述一個有意義的研究問題、調查系統的功能與需求、瞭解系統建置的流程與程序，及研讀相關的文獻以瞭解新的方法與觀念。

二、　發展系統架構：包含發展獨特的系統架構設計，及定義系統元件的功能與它們之間的相互關係。

三、　分析設計系統：包含設計資料庫/知識庫的架構與實現系統功能的流程，及提出幾個系統發展的解決方案並從其中擇一實行。

四、　建置（雛型）系統：包含深入瞭解整個系統建置流程的觀念、架構、設計及系統的所面臨的問題與複雜度。

五、　觀察評估系統：藉由電腦模擬法評估系統，包含評估系統方法之性能，電腦計算之成本等等。

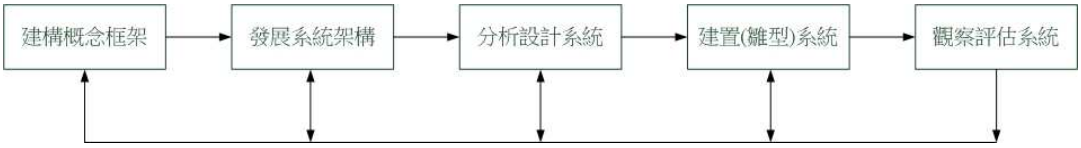

圖 2 系統發展研究流程

資料來源： (Nunamaker、Chen、Purdin, 1991)

系統發展研究法首先定義問題，接著嘗試以發展（雛型）系統的方式解決問題，分析雛型系統架構並逐步實作系統，最後觀察並評估系統以獲得期望的結論並提出建議，同時要採用這個研究方法有五個很重要的準則：

一、　透過系統之建立，目的是要去研究在資訊系統領域一個重要的現象。

二、　研究的結果能對其專業領域有顯著的貢獻。

三、　對於系統所闡明的目標和要求必須是能被測試的。

四、　新的系統必須要較已存在的系統提供更好的問題解決方法。

五、　在建置系統時所獲得的經驗及專業知識能被歸納出來做為日後的使用。

本研究將依據上述步驟與流程，先依據相關文獻搜集與探討，發展各系統元件之功能，如網路爬蟲，中文斷詞系統、LIBSVM 等等，在分析各系統元件後設計系統架構圖，接著依照系統架構圖逐步實作出完整系統，進行系統實驗並觀察實驗結果，最後歸納評估系統。

肆、　雛型系統

本研究參考 Bing Liu (2004)與王正豪 (2010)之研究提出以下系統發展之前提：具有主觀情緒性質之評論行為語句，被評論的屬性詞與意見片語的各項特徵，與其文字距離有一定程度之正相關。依據上述假設與參考楊昌樺等人 (2007)的模式，本研究所建置之雛型系統架構分為五個部份，如圖 3 所示。有關架構所開發之雛型系統的運作流程如圖 4 所示。

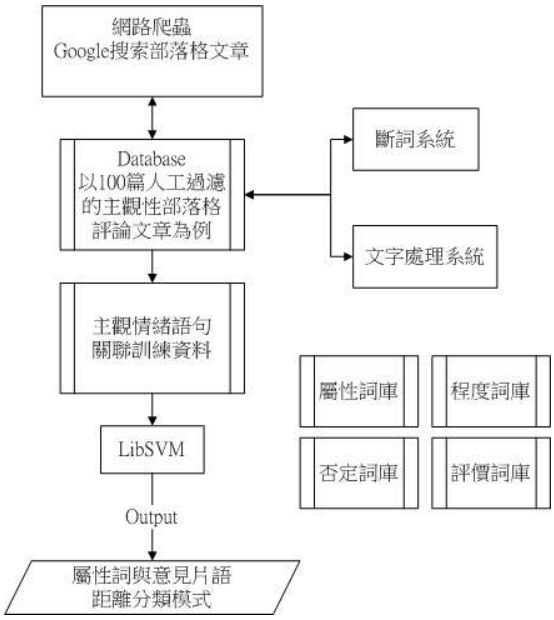

圖 3 本研究部落格文章情感分析系統架構圖

一、　蒐集部落格文章：以網路爬蟲配合 Google 搜尋的結果，取得部落格文章，同時將各項關鍵資訊分析出，如文章本文、人氣指數、發文時間等等。

二、　標示主觀性文章：以人工方式標示出為文章作者本身進行評論的文章，將新聞、廣告、轉錄等文章過濾。

三、　斷詞系統：使用中研院的 CKIP 中文斷詞系統，將文章做初步斷詞，提供與評價詞庫、程度詞庫、否定詞庫、屬性詞庫，目標關鍵字比對的依據。本研究將目標關鍵字也視為屬性詞庫當中的屬性詞。

四、　文字處理系統：分析比對主觀情緒語句的訓練資料，將文章資料正規化，轉換成可供 LIBSVM 的訓練資料或實驗資料。

五、　LIBSVM：將格式化資料進行訓練，產出訓練資料的最佳模式。

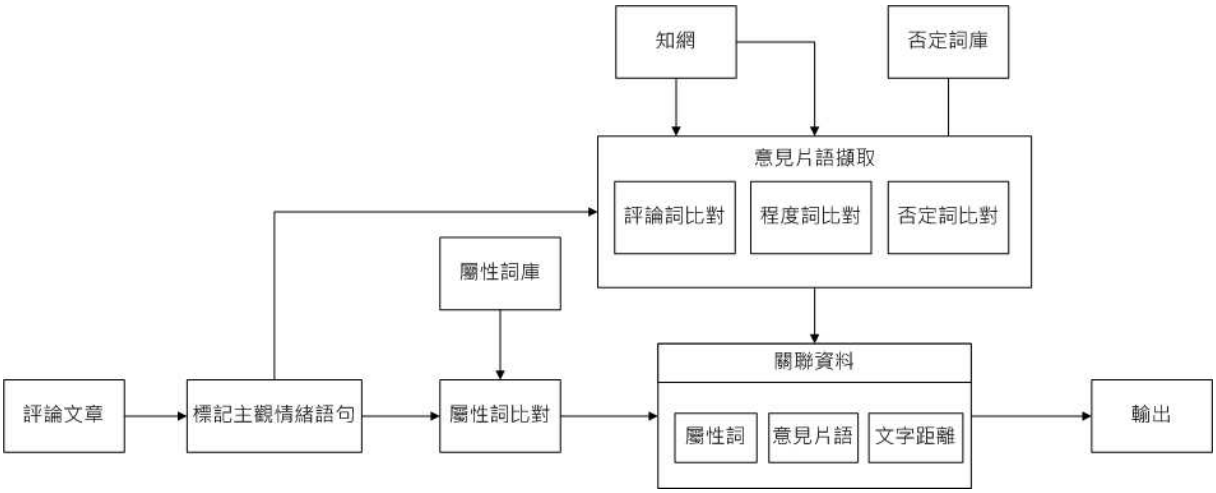

圖 4 本研究雛型系統運作流程

本研究首先透過網路爬蟲，以 3C 產品『iPhone4』為例進行部落格相關文章的擷取，並輔以『開箱』、『使用心得』等關鍵字進行搜尋，再以人工方式將具有主觀情緒評論性質的文章標記出來做為實驗的評論文章，並為了評論文章的客觀性，使用 3 名人員進行評論文章的判讀，需全體判讀為具有主觀情緒評論性質的文章，才可視為本研究實驗的評論文章，本研究以 100 篇之主觀情緒評論文章做為實驗文章。因本研究選擇在語句層級上做探討，故仍然是以 3 名人員從實驗文章中進行主觀情緒語句識別標記，從每篇文章中標記出多個具有主觀情緒評論的語句之開始與結尾，每一語句均含有一至多個的評論行為，並需全體判讀一致才可視為本研究的實驗語句，本研究共標記了 341 句主觀情緒語句作為實驗語句。

Bing Liu 在其研究中定義了評論模型，每個被評論的物件有其所屬的一組被評價的屬性，故本研究以人工方式列出目標關鍵字可能被評價的屬性，如「電池」、「外型」等，共列出 67 項屬性詞做為屬性詞庫 (2008)。再將屬性詞庫與所有主觀情緒語句的斷詞結果做比對，若屬性詞與某一字詞完全相同，則視為該主觀情緒語句的屬性詞，再標記出該屬性詞於該主觀情緒語句的位置。

本研究引用王正豪教授之中文部落格文章意見分析研究 (2010)，將依照其擷取「意見片語」的方法，並將意見片語做為 SVM 分類的各項特徵值。而本研究主要為研究屬性詞與意見片語之距離關係，不探討意見片語之擷取準確率，僅將意見片語視為情感判斷的依據。

本研究使用知網發布的中文情感分析用詞語集，將情感詞子類別詞語集和評價詞子類別詞語集做為評價詞庫的來源，程度詞子類別詞語集做為程度詞庫的來源，並參考其歸納的否定詞清單做為否定詞庫。接著將評價詞庫、程度詞庫、否定詞庫與主觀情緒語句比對，擷取出主觀情緒語句當中的意見片語和意見片語位置。

本研究將屬性詞、意見片語和兩者之間的字詞距離做為一組關聯資料，共擷取出 868 筆關聯資料，提供相關資料輸出至 SVM 做為訓練和測試。

伍、 實驗設計

為有效驗證雛型系統之正確率與效率，本研究設計實驗環境，以實際網路部落格文章為基礎。本研究依據知網所發佈的中文情感分析用詞語集，定義每筆關聯資料中的各特徵如表 1 所示。表 1 中：

編號 1 為屬性詞與評價詞之距離，各項特徵之距離計算均是以斷詞系統的斷詞結果為準。

編號 2 為依據中文情感分析用詞語集區分的正反面情感分類。若屬正面則資料為 1，反面則資料為-1。

編號 3 為評價詞的位置在屬性詞的位置前或後，若屬後面則資料為 1，屬前面則資料為-1。

編號 4 為該意見片語中，是否有程度詞，若有則資料為 1，無則資料為-1。

編號 5 為依據中文情感分析用詞語集的程度詞 6 個等級編號。若編號 4 屬無程度詞，則資料為 0。

編號 6 為程度詞的位置在評價詞的位置前或後，若屬後面則資料為 1，屬前面則資料為-1，若編號 4 屬無程度詞，則資料為 0。

編號 7 為程度詞與評價詞之距離，若編號 4 屬無程度詞，則資料為 0。

編號 8 為該意見片語中，是否有否定詞，若有則資料為 1，若無則資料為-1。

編號 9 為否定詞的位置在評價詞的位置前或後，若屬後面則資料為 1，屬前面則資料為-1，若編號 8 屬無否定詞，則資料為 0。

編號 10 為否定詞與評價詞之距離，若編號 8 屬無否定詞，則資料為 0。

表 1 關聯資料特徵表

編號	特徵名稱	資料範圍
1	屬性詞與評價詞距離	>0
2	評價詞情感正反面	1,-1
3	屬性詞位置相對於評價詞前後	1,-1
4	有無程度詞	1,-1
5	程度詞分級	0-6
6	程度詞位置相對於評價詞前後	1,0,-1
7	程度詞與評價詞距離	>=0
8	有無否定詞	1,-1
9	否定詞箱對於評價詞前後	1,0,-1
10	否定詞與評價詞距離	>=0

本研究觀察 868 筆關聯資料中，屬性詞與評價詞之相對位置分布，大多位於前後 10 個字詞之內，故本研究將資料範圍限定在屬性詞與評價詞前後距離 10 個單位以內的關聯資料，避免距離過大的關聯資料干擾 SVM 分類之準確率。

本研究以 LIBSVM 為工具，將關聯資料特徵中的編號 1 屬性詞與評價詞距離做為分類標籤，其他關聯資料的特徵做為索引，特徵資料做為數值，使用 SVM 訓練出主觀情緒語句的距離分類模式。期望此距離分類模式能有效區分主觀情緒語句與非主觀情緒語句，從而判定文章的主觀評論成分。

有關本研究實驗流程如圖 6 所示。首先將關聯資料隨機區分為訓練資料與測試資料，再將訓練資料轉換為可供 SVM 訓練的格式。SVM 的訓練格式如圖 5 所示。

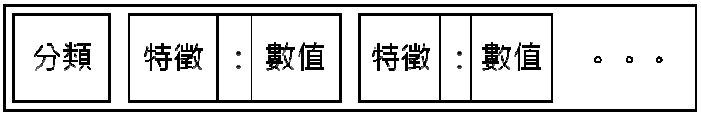

圖 5 SVM 訓練格式

將訓練資料使用 SVM 訓練，並使用 LIBSVM 工具內的 grid.py 工具，自動尋找該訓練資料的最佳參數 c (Cost)和 g(Gamma)。取得測試資料以與訓練資料同樣的格式輸出，交由 SVM 做分類，將分類結果記錄下。接著將測試關聯資料中的結果與分類結果比對，計算準確率。

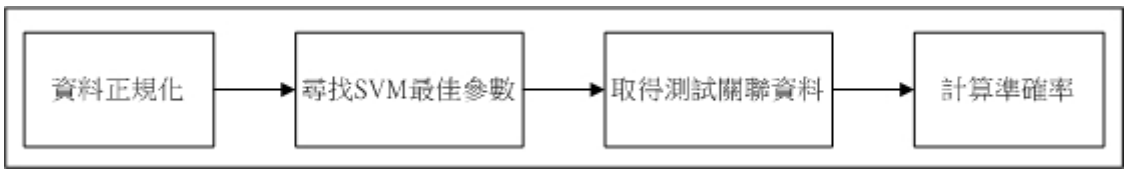

圖 6 本研究實驗流程

為有效驗證本研究雛型系統之正確性與效率，本研究設計四個實驗情境，以下描述實驗進行的過程與分析。

實驗 1. 精確率分析

實驗首先隨機取得限定距離於前後 10 字詞之內的關聯資料，且在每個相同字詞距離的資料中，各取隨機 12 筆資料，使得每一距離的關聯資料數量均相同，總得出 240 筆關聯資料做為 SVM 訓練資料。

使用 LIBSVM 之 grid.py 工具，找出該組訓練資料的最佳參數 c 和 g，使用 SVM 進行分類，輸出該組訓練資料的分類模式。

接著依據該分類模式，隨機取得約 20%的關聯資料做為測試資料，使用 SVM 工具進行分類，將分類結果與測試關聯資料距離做比對，若相同則視為分類正確，不同則視為分類錯誤。最後得出該次訓練的準確率。

從取得訓練資料步驟到最後計算準確率步驟，重覆進行 10 次，各次準確率結果如表 2 所示，轉換以圖表顯示如圖 7。

表 2 限定距離於前後 10 字詞之準確率

項次	1	2	3	4	5	6	7	8	9	10
準確率	20.16%	17.72%	18.53%	26.38%	24.62%	20.46%	18.07%	19.54%	18.96%	19.10%

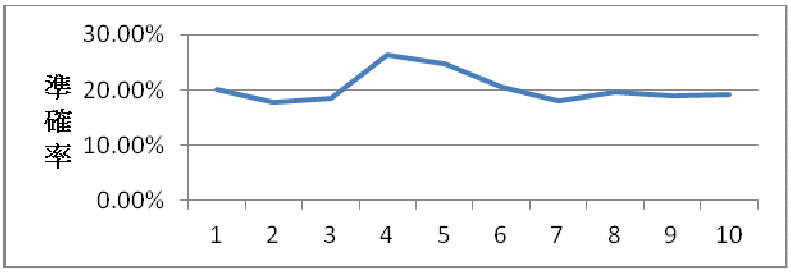

圖 7 限定距離於前後 10 字詞之準確率折線圖

本次實驗結果，平均準確率為 20.35%，顯示在文字距離 10 個分類當中，剛好分在正確分類上的分類效果有些許程度的集中，顯示文字距離與關聯資料全部屬性有些許的正相關。

實驗 2. 容許誤差精確率

本次實驗依據 Bing Liu (2004)的研究，其研究中曾經使用計算固定文字距離之內情緒詞的方法，來計算某屬性詞的情緒相關性。故在本次實驗中，重覆實驗 1 的流程，僅在比對方法上做了些許修改，將分類結果與測試關聯資料距離比對的方法，從分類完全正確修改成在容許誤差內的分類距離即視為分類正確。本次實驗的誤差範圍從 1 至 10 做各誤差距離的準確率計算。有關 1 至 10 各誤差距離的準確率數據如表 3 所示，以平均準確率轉換以圖表顯示如圖 8 所示。

表 3 誤差範圍從 1 至 10 做各誤差距離之準確率

誤差距離	0	1	2	3	4	5	6	7	8	9	10
平均準確率	17.36%	37.98%	53.88%	65.44%	73.49%	83.74%	91.55%	95.90%	99.29%	100%	100%

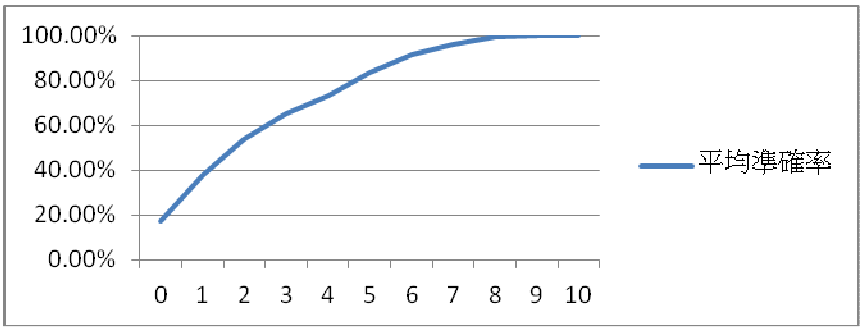

圖 8 誤差範圍從 1 至 10 做各誤差距離的準確率折線圖

依據本次實驗結果，容許誤差距離 3 時，準確率可大幅提升至 65.44%，而相關文獻中，相同研究目標之準確率約為 50%，顯示在分類結果的前後距離 3 個字詞內，可有效預估測試語句的屬性詞與意見片語的距離，從而判定該測試語句的主觀情緒性質。

觀察誤差距離與平均準確率的分布，可看出誤差距離約為 3 以上時，準確率呈現些許的緩和提升，表示準確率僅需計算誤差範圍 3 之內的關聯資料，即可達到一定程度之準確率提升，而不必計算 4 至 10 的誤差距離，以此降低計算誤差距離的運算成本。

實驗 3. 特徵分類相關程度

本實驗為得知列出之關聯資料屬性，與分類準確率之對應關係，將各屬性分類為 3 類屬性，分別為評價詞類屬性、程度詞類屬性和否定詞類屬性(如表 4)，各關聯資料屬性分詞類如下表 4，分別將各分類屬性進行準確率的誤差準確率運算，如下表 5 所示，轉換以圖表顯示如圖 9 所示。

表 4 屬性分類表

編號	屬性名稱	屬性詞類
1	屬性詞與評價詞距離	評價詞類
2	評價詞情感正反面	評價詞類
3	屬性詞位置相對於評價詞前後	評價詞類
4	有無程度詞	程度詞類

5	程度詞分級	程度詞類
6	程度詞位置相對於評價詞前後	程度詞類
7	程度詞與評價詞距離	程度詞類
8	有無否定詞	否定詞類
9	否定詞箱對於評價詞前後	否定詞類
10	否定詞與評價詞距離	否定詞類

表 5 各分類屬性之誤差準確率

各誤差距離	全部屬性	評價詞類	程度詞類	否定詞類
1	37.987%	18.372%	35.5938%	24.3726%
2	53.8865%	29.0154%	49.7776%	36.0163%
3	65.4485%	39.5048%	60.6436%	47.6027%
4	73.4995%	49.7552%	71.1707%	60.8804%
5	83.749%	60.5307%	80.7551%	70.1589%
6	91.5596%	68.8601	87.062%	76.1259%
7	95.9075%	78.763%	94.0409%	84.3616%
8	99.2926%	91.1157%	98.2657%	94.0779%
9	100%	100%	100%	100%
10	100%	100%	100%	100%

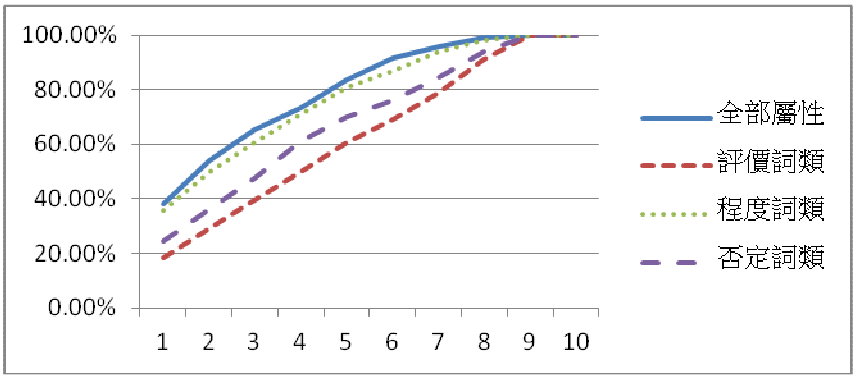

圖 9 各特徵詞類的準確率

由圖 9 中各屬性詞類的準確率變化，可觀察到程度詞類的變化較為貼近全部詞類的誤差準確率變化，顯示關聯訓練資料中，程度詞類的出現較能提升各誤差準確率。其因素可能為程度詞根據知網的分類，將程度詞分為 6 等級，與其他屬性的資料範圍相比，範圍較廣故較能區分各主觀情緒語句與距離的關係。

實驗 4. 運算成本比較

本實驗為計算 SVM 訓練出的距離分類模式，若使用距離分類出的距離做距離比對方法，與未使用距離分類做距離比對的方法，兩者的運算成本比較，以下分別敘述兩者之比對方法。

一、 未使用距離分類算法

屬性詞與評價詞之比對方式將從每筆關聯資料中的屬性詞絕對位置起算，計算固定

距離範圍內的詞，比對該詞是否為評價詞，比對一次則視為一次的運算成本，本研究已將範圍限定在距離 10 個單位以內的詞，而距離範圍內的比對順序將是從距離 1 個單位的詞開始，依序到距離 10 個單位的詞，如圖 10 所示。

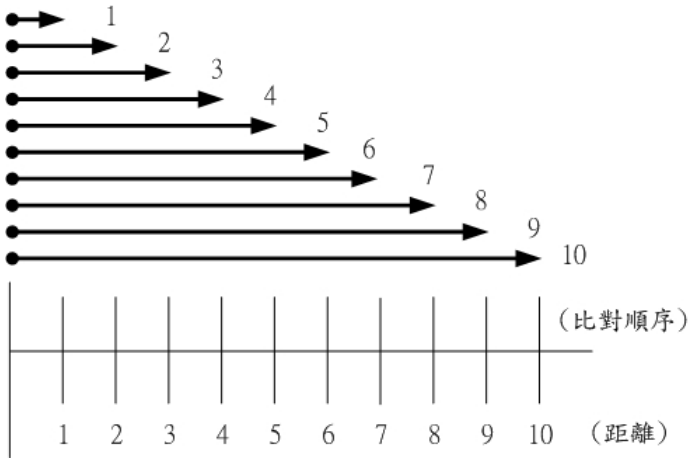

圖 10 未使用距離分類算法之比對順序

二、 SVM 距離分類算法

屬性詞與評價詞之比對方是從每筆關聯資料中的屬性詞絕對位置起算，使用經過 SVM 分類出之距離，計算該距離誤差範圍 3 以內的詞，比對該詞是否為評價詞，比對一次則視為一次的運算成本，而比對順序將是從分類距離開始，由誤差範圍 1 個單位依序到 3 個單位結束。比對順序如圖 11 所示，x 表示經 SVM 分類之距離。

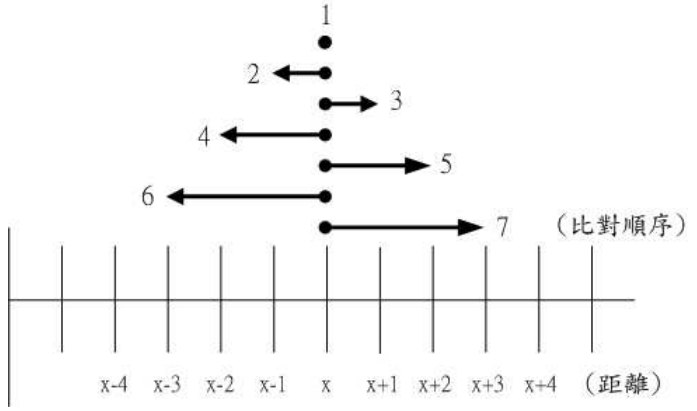

圖 11 使用 SVM 距離分類算法之比對順序

實驗首先使用 SVM 輸出該訓練資料的分類模式，接著取得約 50%的測試資料，分別計算以上兩種比對方法的運算成本，重覆進行 1000 次。

實驗結果為方法二的運算成本均少於方法一的運算成本，兩種方法的運算成本變化如圖 12 所示，由圖 12 可看出兩種方法的運算成本均呈現線性成長，但成長程度略為不同，方法二的程度較低，計算各次實驗的運算成本差異，方法二的運算成本平均可為方法一減少 21.75%的運算成本。

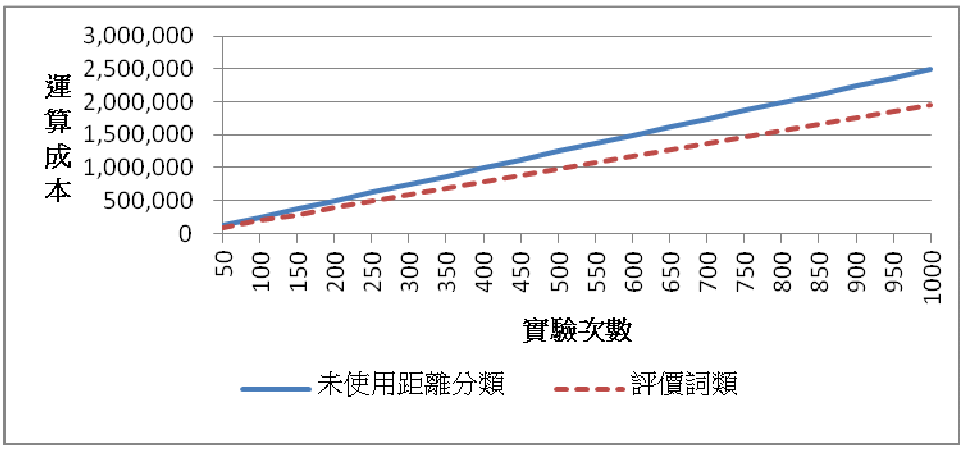

圖 12 運算成本變化比較圖

陸、 結論及未來發展

本研究透過人為的方式蒐集主觀情緒文章，並於文章中標記主觀情緒語句，在語句層級上，嘗試在大量的主觀情緒語句資料中，能歸類出主觀情緒語句的評價模式，以成為區分主觀情緒語句與非主觀情緒語句的規則。引用相關文獻中的擷取意見片語方法，成為情感判斷的依據，建立評價詞庫、程度詞庫、否定詞庫，再以「iPhone4」為例成為目標關鍵字，列出其相關的屬性詞庫。再將屬性詞庫與意見片語包含距離、程度詞分級等等的相關屬性，成為 LIBSVM 工具的 SVM 訓練資料與測試資料，歸納出主觀情緒語句在文字距離上的分類模式。透過關聯資料的各個屬性欄位，歸納出該筆關聯資料若符合主觀情緒語句特徵，其正確的文字距離應為多少，再依此與實際文字距離做比對來判定該筆關聯資料是否為主觀情緒語句。

從實驗中發現，若使用 SVM 歸納的分類模式來分類主觀情緒語句當中屬性詞與文字的距離，完全正確分類僅有約 20%的準確率，但若將容許誤差距離設為 3，其分類正確率可提升至 65%左右。對照以往以文字距離為主的情感分析方法，其距離的計算範圍需先定義出一個固定的距離再作範圍內的運算，若以本研究之方法做距離分類運算，可直接使用 SVM 判讀之距離進行誤差內範圍運算，在可達到相當的準確率前提下，可有效降低計算誤差距離的運算成本約 21.75%。

在未來的研究中，應朝提高距離分類準確率為目標，修正本研究分類的各項依據，如意見片語、關聯資料屬性等等，可能包括將以修辭學的角度來比對主觀情緒語句做為訓練分類的屬性，在距離分類上可使 SVM 分類更精確。

參考文獻

1. 王正豪、李啟菁，《中文部落格文章之意見分析》，碩士論文，國立台北科技大學資訊工程研究所，2010。
2. 林晏僖、高照明、高成炎，《中文名詞組的辨識：監督式與半監督式學習法的實驗》，2008 自然語言與語音處理研討會，頁 180-193，台北，2010。
3. 林智仁，＜Welcome to Chih-Jen's Lin's Home Page＞，網址：

http://www.csie.ntu.edu.tw/~cjlin/index.html，上網日期：2012 年 2 月 1 日。
4. 孫瑛澤、陳建良、劉峻杰、劉昭麟、蘇豐文，〈中文短句之情緒分類〉，2010 自然語言與語音處理研討會，頁 184-198，暨南大學，2010。
5. 高照明，〈中文詞彙語意資料的整合與擷取：詞彙語意學的觀點〉，2007 自然語言與語音處理研討會，頁 257-272，台北，2010。
6. 黃文奇、吳世弘、陳良圃、谷圳，〈中文文字蘊涵系統之特徵分析〉，2011 自然語言與語音處理研討會，頁 281-296，台北，2011。
7. 楊昌樺、陳信希，〈以部落格文本進行情緒分類之研究〉，2006 自然語言與語音處理研討會，交通大學，2006。
8. 楊昌樺、高虹安、陳信希，〈以部落格語料進行情緒趨勢分析〉，2007 自然語言語音處理研討會，頁 205-218，台北，2007。
9. Hu, M., and Liu, B., "Mining Opinion Features in Customer Reviews," Proceedings of the 19th National Conference on Artificial Intelligence, 2004, pp. 755-76.
10. Hu, M., and Liu, B., "Mining and Summarizing Customer Reviews," Proceedings of he 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD-2004), 2004, pp. 168–174.
11. Jindal, N., and Liu, B., "Mining Comparative Sentences and Relations," AAAI., 2006, pp. 1331-1336.
12. Ku, L. W., and Chen, H.H., "Mining Opinions from the Web Beyond," Journal of American Society for Information Science and Technology, 2007, pp. 1838-1850.
13. Liu, B., *Encyclopedia of Database Systems*, 2004.
14. Liu, B., Sentiment Analysis and Subjectivity. Natural Language Processing, 2010.
15. Liu, B., "Opinion Observer: Analyzing and Comparing Opinions on the Web," In Proceedings of the 14th international Conference on World Wide Web, Japan, 2005, pp. 342-351.
16. Nunamaker, J. R., Chen, J. F., and Purdin, T. D. M., "Systems Development in Information Systems Research," Journal of Management Information Systems, Vol. 7, 1991, pp. 89-106.
17. Ounis, I., Rijke, M., Macdonald, C., Mishne, G., and Soboroff, I., "Overview of the TREC-2006 Blog Track," In Proceedings of TREC-2006, USA, 2007.
18. Pang, B., and Lee, L., "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," In Proc. 42nd ACL, Spain, 2004, pp. 271-278.
19. Taku, K., and Yuki, M., "Use of Support Vector Learning for Chunk Identification," Proceeding of CoNLL-2000, Lisbon, 2000, pp. 142-144.
20. Vapnik,N. V., *The Nature of Statistical Learning Theory*. Springer, 1995.
21. Wan, X. J., "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis," In Proceedings of Empirical Methods in Natural Language Processing, 2008, pp. 553-561.

A study of sentimental analysis in blog articles

Ruey-Shiang Shaw
Department of Information Management, Tamkang University
rsshaw@mail.tku.edu.tw

Qing-Shan Jiang
Shenzhen Institute of Advanced Technology Research Institute
qs.jiang@siat.ac.cn

Chin-Feng Tsao
Technical Seed Corp.
gino@t-seed.com

Chih-Wen Chien
Department of Information Management, Tamkang University
fresh.chien@mail.im.tku.edu.tw

Abstract

How to capture the emotional opinion efficiently and quickly in a lot of network community articles is an important basic work for Sentiment Analysis. This study is discussed in the sentence-level analysis of the subjective emotional comment sentences in the article, and try to find out a judgment model of subjective emotional sentences and non-subjective emotional sentences. This study adopts Systems Development Methodology, and it uses SVM tool to train and test the subjective emotional sentences. In this experiment, it compares the classification of SVM with the actual classification and it calculates the accuracy as the basis for system verification.

On the basis of the classification of SVM, this study found if the tolerance value of the distance between attribute word and opinion phrase is set to 3, it will significantly improve the accuracy of distance classification and reduce the computing cost of distance comparison, and then the classification feature of the attribute words is better than negative words and opinion words in recognizing the distance relationship between attribute word and opinion phrase.

Keywords: Sentiment analysis, SVM, HowNet, Subjective emotion