以部落格語料進行情緒趨勢分析

楊昌樺　高虹安　陳信希

國立台灣大學資訊工程學系

{d91013, r95116, hhchen}@csie.ntu.edu.tw

摘要

部落格提供大量具有時間標記的文本，為語言處理所需豐富語料來源。本文針對文本的時間標記特性，將其切分成不同時間域(Time Domain)的語料子集合，綜合個別時間域所提供的語料，觀察目標觀點(sentiment，通常包含意見與情緒)在橫跨時間域的變化，作為觀點趨勢分析的基礎。為了獲得不同時間域的部落格文本，本研究提出部落格資訊系統，收集跨時間域的文本。同時以情緒分析為例，以特定查詢在部落格資訊系統反饋的相關文本，獲得各時間域的情緒特徵，藉以解讀網路空間人們對特定議題所反應的情緒變化。

1. 緒論

近年來，全球資訊網(World Wide Web; Web)上各種資訊系統的發明，持續改變人們對資訊吸收與處理的方式。以 2005 年作為一個概括的分水嶺，分析使用者的習性，可發現：2005 年之前，人們從接觸網路，到逐漸習慣閱讀 Web 上的內容，包含新聞報導、旅遊情報、生活資訊、投資消息、工作機會等。2005 年之後，人們開始廣泛地創造 Web 上的內容，包含製作自己部落格、相簿遊記、影音紀錄等。Web 服務提供者，為了能滿足前者的閱讀需求，激發了內容網站(如 nyyimes.com、cnn.com)、與服務網站(包括入口網站，如 Yahoo![1]，及搜尋網站，如 Google[2])的興起。而為了能涵蓋後者的創造需求，也帶

[1] http://www.yahoo.com/

[2] http://www.google.com/

動了部落格網站(如 Blogger[3])、相片網站(如無名小站[4])、影音網站(如 YouTube[5])等的蓬勃發展。以媒體的觀點分析之，前者透過特定企業將 Web 視為大量資訊的媒介，提供的站台延續大眾媒體的角色。後者藉由 Web 使用者社群參與，創造新式資訊發佈型態，通稱為社群媒體。

近來社群媒體所創造的資源吸引很多學者的注意，本文針對部落格(或稱網路日誌、Weblog、Blog)所提供的文本，進行語言處理方面的探討。部落格系統提供簡單的介面，讓使用者發表具時間標記的文章，因此有越來越多的人們開始使用部落格在網路上分享每天的生活經驗、發表對事物的看法與心情。根據部落格搜尋引擎Technorati[6]的報告指出，全球部落格的數量已超過7,000萬個，並且平均每天有超過12萬個部落格成立，因此整個部落格空間(Blogosphere)每天所能貢獻出的新文本更在此數量之上。這份報告中同時也指出，目前部落格空間以日文及英文使用者居多，各佔37%及36%，而中文目前所佔比例是8%，但有增長的趨勢，本文即以中文部落格文本為主要的研究對象。

在社群媒體的框架下，人們在使用部落格搜尋引擎時，不但想找到較專業或具代表性的部落格，同時也想找到一般使用者所提出的心得及想法。TREC自2006年開始舉辦Blog Track[7](Macdonald, de Rijke, Mishne, and Soboroff, 2006)，其競賽項目說明了使用者上述的資訊需求。其中Opinion Retrieval Task是針對特定議題找出使用者表達意見的文章，並判斷該意見文章的正負面傾向。另外，Blog Distillation (Feed Search) Task是找出持續對某特定議題關注的部落格。舉例來說，使用者可能想在一個著名的歌唱大賽結束後，瀏覽各部落格最新發表的相關文章，並挑有興趣的閱讀。使用者也可能剛接觸古典音樂，想找專門討論古典音樂的部落格，並在肯定某部落格的豐富內容後，持續訂閱該部落格。

---

[3] http://www.blogger.com/

[4] http://www.wretch.cc/

[5] http://www.youtube.com/

[6] http://www.sifry.com/alerts/archives/000493.html

[7] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

相較於從Web語料判斷使用者的意見和情緒(Ku and Chen, 2007; Lin, Yang, and Chen, 2007)，從部落格文本中挖掘使用者觀點有如下的好處：個人化導向和時間性。部落格是個人在網路世界最直接的發聲工具，使用者經認證登錄後，就可以特定的虛擬身分進行發布文本的行為。相較之下，Web是零零總總資訊的巨大集合，所提供的各式文本在結構或型態上都可能不一致。當Web網頁一經發布後，除非進行更動或刪除，就會一直存在於網路空間裡，儘管能透過搜尋引擎取回文本進行判斷，也較無機制獲得網頁發布或更動的時間。相較之下，部落格文本通常具有時間標記，在呈現上也是依照時間從最近到最舊排列，因此透過部落格收集使用者觀點，除了能根據所有的文本進行分析外，也有機會分析出不同時間域(如最近、上周、去年)使用者的特徵。

儘管由如上所述的特點，透過部落格擷取使用者觀點，仍存在一些語料分析時必須考量的議題，例如嚴(2007)曾指出，部落格文本中有一半的文章是來自轉錄，而不是部落格作者自己撰寫。這些轉錄文章的內容大部份是從新聞網站、或一般的官方網站，經由複製、轉貼到使用者自己部落格上。這是因為在便利的網路環境下，人們很容易取得其他資訊，也很方便再把資訊傳播出去，因此傳播出去的不見得就是該使用者本身的立場或意見。在此情況下，系統可能收集到不同於使用者觀點的雜訊。

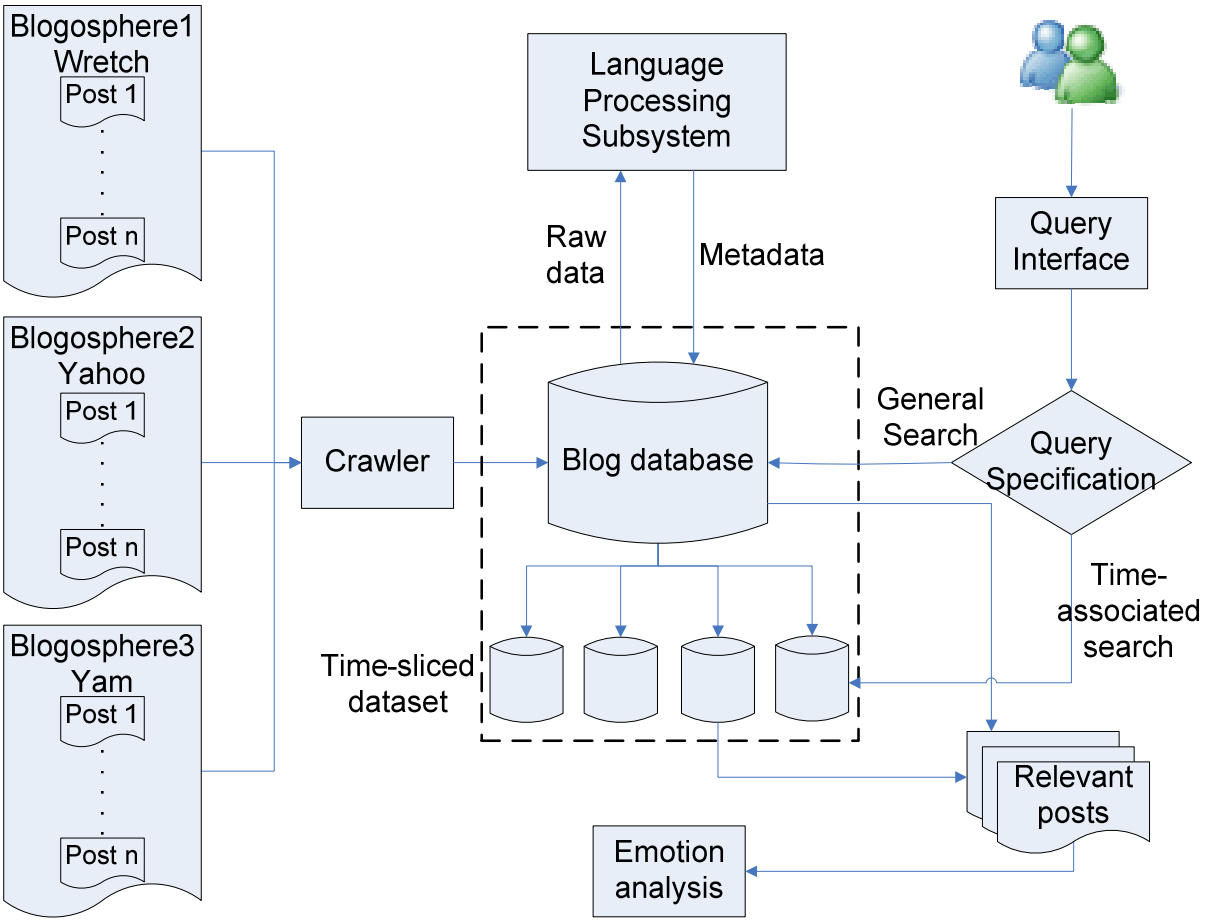
雖然透過部落格容易達成溝通，但相對來說，也較容易產生衝突。由於在部落格世界中，人們的身分通常是由一個虛擬的ID或暱稱所代表，因此在與社群意見不一致時，較容易產生加油添醋、謾罵，甚至言語攻擊等行為。倡議Web 2.0的Tim O'Reilly在2007年曾提出一個「部落客行為守則」 (Blogger's Code of Conduct[8])，希望部落客能自我維護言論自由的環境。然而從語料分析系統的觀點，無論是好或是壞的言論，都是一種文本的呈現，會一視同仁地收集回來，但在日後分析處理時，是否能過濾掉部落客不良動機所帶來的雜訊，也是一個需要重視的議題。

本文嘗試在不同時間域上對部落客情緒進行分析，建置部落格資訊系統，以獲取具

[8] http://radar.oreilly.com/archives/2007/03/call_for_a_blog_1.html

時間特性的文本語料。內容安排如下：第二節列出系統架構，第三、四節針對系統兩個重要核心－文本收集與整合、情緒分析與趨勢－加以說明，第五節是舉出一些應用範例，並在第六節提出結論。

2. 部落格資訊系統架構



圖一、部落格情緒趨勢分析系統架構圖

本研究所探討的部落格情緒趨勢分析系統包括四部分：收集與整合部落格文本資訊、資料庫與語言處理子系統、使用者查詢流程、以及情緒分析模組。第一部分及第四部分將於第三、四節再作詳細探討。

關於資料庫與語言處理子系統部分，在 Crawler 經由查訪部落格文本單位後，會先剖

析出部落格欄位資訊，將其存放在對應的部落格資料庫(Blog Database)。另外，各欄位所包含的語言資訊，如須經斷詞處理以獲得進一步索引時，我們的語言處理子系統(Language Processing System)會利用 Stanford Natural Language Processing Group[9] 所開發的 Stanford Chinese Word Segmenter[10]加以處理，並回傳至原資料庫，以 Metadata 的型式協助往後搜尋的進行。

部落格資料庫本身可透過文本的時間資訊，在實體資料集(Physical Dataset)上產生不同時間域子集合(Time-sliced Dataset)，上述兩種模式的資料集可支援以下不同的查詢應用：第一種是基本查詢(General 查詢)，例如於使用者介面可以提供搜尋系統，讓使用者鍵入擬查詢的關鍵字後，由系統回傳相關的部落格文本資訊，使用者可以瀏覽分析部落格文本資訊後，透過永久網址(Permalink)鏈結到該部落格文本原先的網頁。第二種資料集可支援時間相關的查詢(Time-associated Search)，例如透過關鍵字查詢相關文本在不同時間域的情緒呈現趨勢，系統可依照時間域與情緒分類繪製情緒波動的列表或趨勢圖。

3. 部落格文本資訊之收集與整合

部落格文本位於不同的部落格伺服器，這些伺服器可能是使用者自行架設、委託代管之程式或機器，或是由企業廠商所提供之服務平台。基於語料平衡性與完整性的考量，一個部落格資訊系統需要能收集不同來源的部落格文本，並能將各式文本整合成為一個資料集合。如同收集 Web 的材料時所面臨到的 Deep Web (He et al., 2007)問題，這裡我們也相對綜合出一個 Deep Blogosphere 的概念，與探索 Deep Blogosphere 的因應之道。

首先在剖析 Blogosphere 的結構後可了解到，每一篇部落格文本都有所謂的 Permalink，一個完整 Deep Blogosphere 的文本探索，簡言之就是能拜訪過世界上所有的 Permalink。然而這些 Permalink 大部分都由各部落格伺服器的資料庫所維護，其網址通常不代表一個真正的實體網頁，而根據其網址所瀏覽到網頁上的鏈結(Hyperlink)，大部分也

---

[9] http://nlp.stanford.edu/

[10] http://nlp.stanford.edu/software/segmenter.shtml

是由各部落格伺服器自動產生，如導引首頁鏈結、圖片鏈結、廣告鏈結、作者資訊鏈結、社群推薦鏈結等。

透過這樣的剖析，可以發現透過傳統 Hyperlink 樹狀拜訪所有網頁結點模式，將難以探索完所有 Deep Blogosphere 的文本。解決之道是在文本之上先維護一個有關所有部落客(Bloggers)的人口普查，意即如果能夠知道世界上所有 Bloggers 的列表，再定時查訪各 Bloggers 所發表的最新文本，以獲得新 Permalink，便能保證在一個實行時間點之後，能夠收集到所有部落格文章。

為了實驗上述概念，本研究將 Deep Blogoshpere 的範圍簡化，將「Yahoo!奇摩部落格」、「無名小站網誌」、「yam 天空部落」視為三個虛擬的 Bloggers，並從六月份開始定期查訪該三個"Bloggers"的最新文章。我們選自 6/20 至 7/8 為止收集到的 322,792 筆部落格文本資訊，供本研究分析。虛擬 Blogger 文章數，及相關統計資料如表一所示。其中「無名小站網誌」每天所能查訪的部落格文本資訊數為最多，接近 1 萬筆，「yam 天空部落」最少，僅約 250 筆。

表一、部落格文本資訊查訪統計

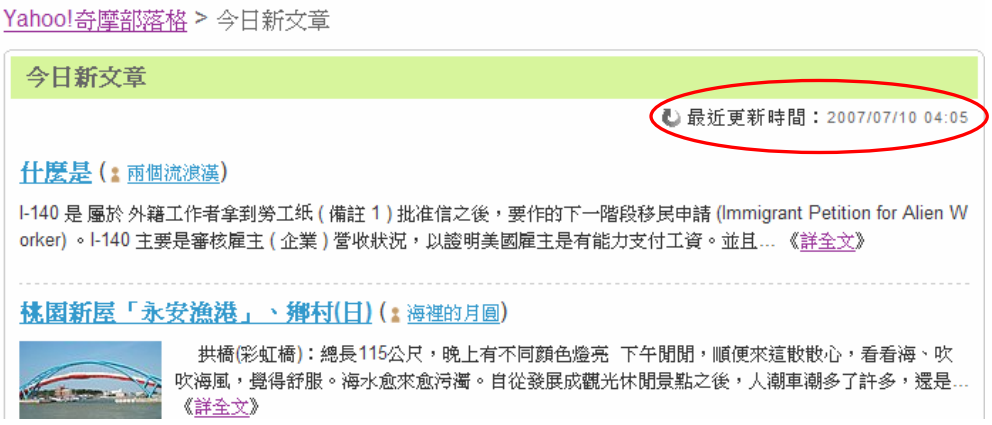
虛擬部落客	文本分析數	每日平均	查訪最新文章方式
Yahoo!奇摩部落格	132,661	6,982	RSS、動態網頁
無名小站網誌	185,234	9,749	Ping Server
yam 天空部落	4,897	258	動態網頁(首頁)

為了查訪Bloggers的最新文本資訊，得因應各部落格伺服器不同的特性，本研究歸納出兩種不同的方式來進行部落格文本資訊的收集，以下分述之。

3.1 查訪部落格網頁列表或 RSS

表一列出各虛擬部落客最新文章查訪方式，以「Yahoo!奇摩部落格」為例，其在首頁「最新文章」區塊提供如圖二紅圈處所標示的RSS按鈕，透過此RSS按鈕可導引到特定的URL[11]，RSS (Resource Description Framework)是部落格空間常用作為資料交換的XML格式規範，本例的URL即以XML的形式列出100篇最新文章。另外該伺服器亦提供「今日新文章」動態網頁[12]，如圖三所示，同樣也可作為查訪部落格文本資訊之依據。然而此網頁較前者不同之處在於部落格文本涵蓋率，以及即時性的比較上。因為其列表列出當天凌晨之前所發布的所有文章，故在部落格文本資訊的涵蓋率上，會較前者完整許多，然而其在即時性上的表現上較前者來得差，原因在於其文章列表僅在當天凌晨更新一次，如圖三紅圈標示處所示。

圖二、Yahoo!奇摩部落格之「最新文章」區塊



圖三、Yahoo!奇摩部落格之「今日新文章」網頁

---

[11] http://tw.blog.yahoo.com/rss/newarticle.xml

[12] http://tw.blog.yahoo.com/newarticle/newarticle.php

類似的方法亦適用於「yam天空部落」。「yam天空部落」首頁[13]即有一區塊列出新發布文章，例如可每隔十分鐘查訪其首頁，對此區塊作剖析以收集部落格文本資訊。

3.2 利用 Ping Server 即時更新文章列表

前一小節所述的RSS規範在部落格空間產生出一種推播(push)的資料流向，因此衍生出一種所謂Ping的資料交換方式，例如當部落客發布一篇新文章時，部落格常會提供自動Ping一個或多個伺服器的服務，亦即發送一個XML-RPC(遠端程序呼叫）信號給一個或多個所謂的 “Ping Servers”。這些“Ping Servers”會藉由收到的信號來產生一個列表，列出有新文本發布或更動的部落格網址。目前開放的Ping Server像是VeriSign公司的Weblogs.com[14]或是Yahoo!公司blo.gs[15]，皆允許網路服務者訂閱其部落格列表，例如部落格搜尋引擎可以藉由查訪最新更新的部落格，來提供使用者較新的搜尋結果。

本研究利用Ping Server的概念，選擇Weblogs.com所提供的“weblog change list”[16]作為部落格文本資訊收集之依據。該列表列出最近五分鐘更新的部落格網址，經由隨機取樣調查發現，中文語料的部落格文本以「無名小站網誌」數量最多，探討原因應為「無名小站網誌」提供自動Ping到Weblogs.com的功能，因此可以預期在透過定時下載此列表的方式，從各「最近五分鐘更新」的部落格文本資訊，累積成從特定時間點之後「所有」的部落格文本資訊。

4. 情緒分析模組

以部落格文本作為語料，Mishne (2005)使用Livejournal[17]標記文本發表時心情的情緒符號，訓練Support Vector Machine (SVM; Cortes and Vapnik, 1995)在文本層次的心情分類

---

[13] http://blog.yam.com/

[14] http://Web.weblogs.com/

[15] http://blo.gs/

[16] http://rpc.weblogs.com/shortChanges.xml

[17] http://www.livejournal.com/

器。Mishne (2006)更進一步在所有觀察的時間域上，以圖型闡釋部落格世界整體的心情指數，例如在情人節前後感受到愛(Love)，在夜間有喝醉(Drunk)的感受。楊和陳(2006)則收集部落格中帶有表情符號的句子，訓練SVM在句子層次的情緒分類器。在以上所述的研究中，部落格文本因在網頁呈現方式上所特別能包含的情緒符號，皆用來當作心情或情緒標記(Taggings)，而文本或文句所包含的關鍵字則形成了所謂的特徵值(Features)。Yang等人(2007)使用了更大規模代有表情符號的部落格語料，進行自動化情緒字典的抽取。本研究參考其情緒字典抽取方式，將文句中具有情緒詞彙的出現作為成特徵值，訓練出包含喜(joy)、怒(angry) 、哀(sad)、樂(happy) 共四類情緒的文句情緒分類器。各情緒分類在正負面傾向和能量波動的程度有所區隔，其區別與相關詞彙如表三所示，如「怒」類除了屬於負面情緒外，其能量波動的程度也較大。四類情緒在文句的呈現上，其中「喜」類情緒出現在感到愛慕、喜好、狂熱的語句中，例如：

「我最喜歡Jolin了!」

「怒」類情緒出現在感到生氣、憤怒、咒罵的文句中，例如：

「可惡!早知道就不要出門了。」

「哀」類情緒出現在感到低潮、痛苦、難過、同情的文句中，例如：

「嗚嗚...可魯真的很可憐」

「喜」類情緒出現在感到高興、開心、有趣的文句中，例如：

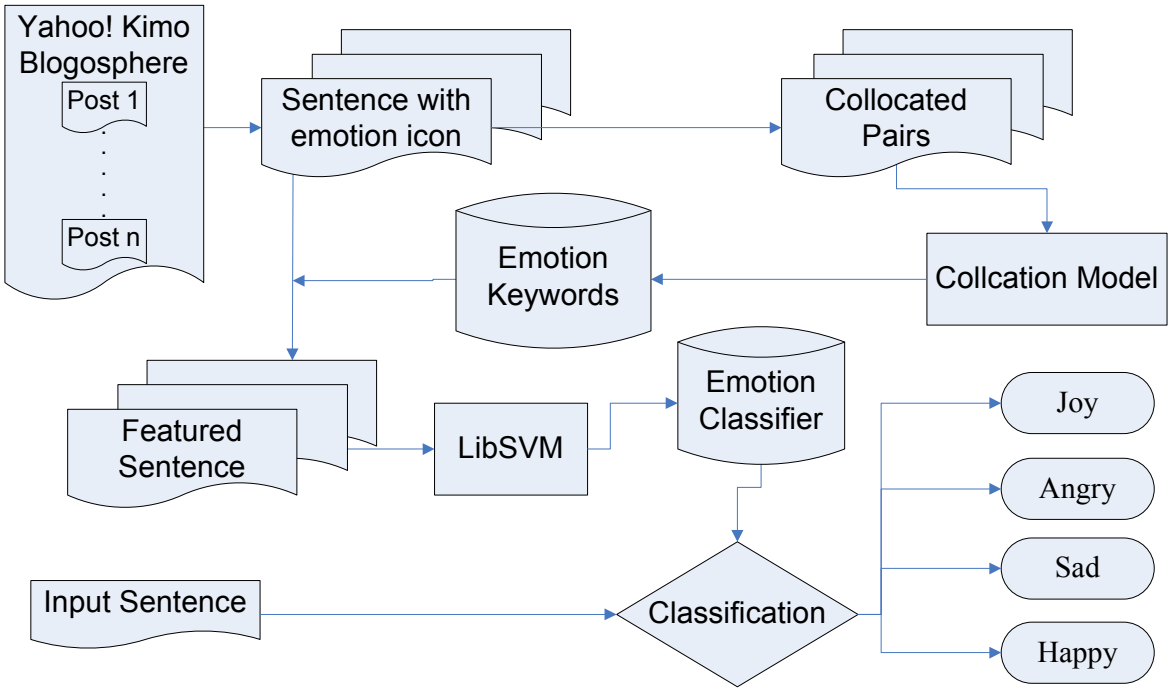
「計畫出遊又碰到好天氣實在很開心。」

表三、情緒分類器訓練時間分析

情緒分類	說明	相關詞彙
喜(joy)	情緒傾向：正面 能量波動：較大	愛、幸福、可愛、喜歡、 謝謝、害羞、感動、寶貝
怒(angry)	情緒傾向：負面 能量波動：較大	生氣、討厭、氣死、可惡、 幹、煩、罵、哼
哀(sad)	情緒傾向：負面 能量波動：較小	哭、痛、嗚、難過、 淚、傷、慘、可憐
樂(happy)	情緒傾向：正面 能量波動：較小	哈哈、開心、好笑、高興、 不錯、加油、很好、好玩

對於情緒的分類與與詞彙的使用有基本的定義後，情緒分析模組的實作流程如圖四所示。首先以Yahoo!奇摩部落格2006年1月到6月所有帶有表情符號的句子作為語料，取出帶有表情符號標記的文句共560,127筆，透過Collocation Model選出500個情緒詞彙。包含該情緒詞彙的文句經由特徵值轉換形成訓練資料，該資料用以訓練情緒分類器，分類器的實作使用Fan等人(2005)所提出的LibSVM[18]工具，以LibSVM工具所訓練的分類器模型則整合入部落格資訊系統，藉以判斷相關文章其構成文句的情緒傾向。

關於訓練資料量部分，於四類情緒各挑選30,000個文句，計12萬句進行情緒分類器的訓練。相較於楊和陳(2006)僅使用約4,000筆訓練語料，本實驗使用了30倍大的訓練集，訓練資料集數量雖然可以繼續嘗試擴大，但是在現今Intel Xeon 5320工作站環境下，30倍大的訓練集已增加了4,120倍的訓練時間。不同訓練資料大小與相對訓練時間如表三所示，當訓練資料僅有4,000筆時，所需的訓練時間僅需6秒，但訓練資料變成兩倍時，所需要的訓練時間竟增加成4倍之多，而本文所採用的分類器則需約7小時才能訓練完成。

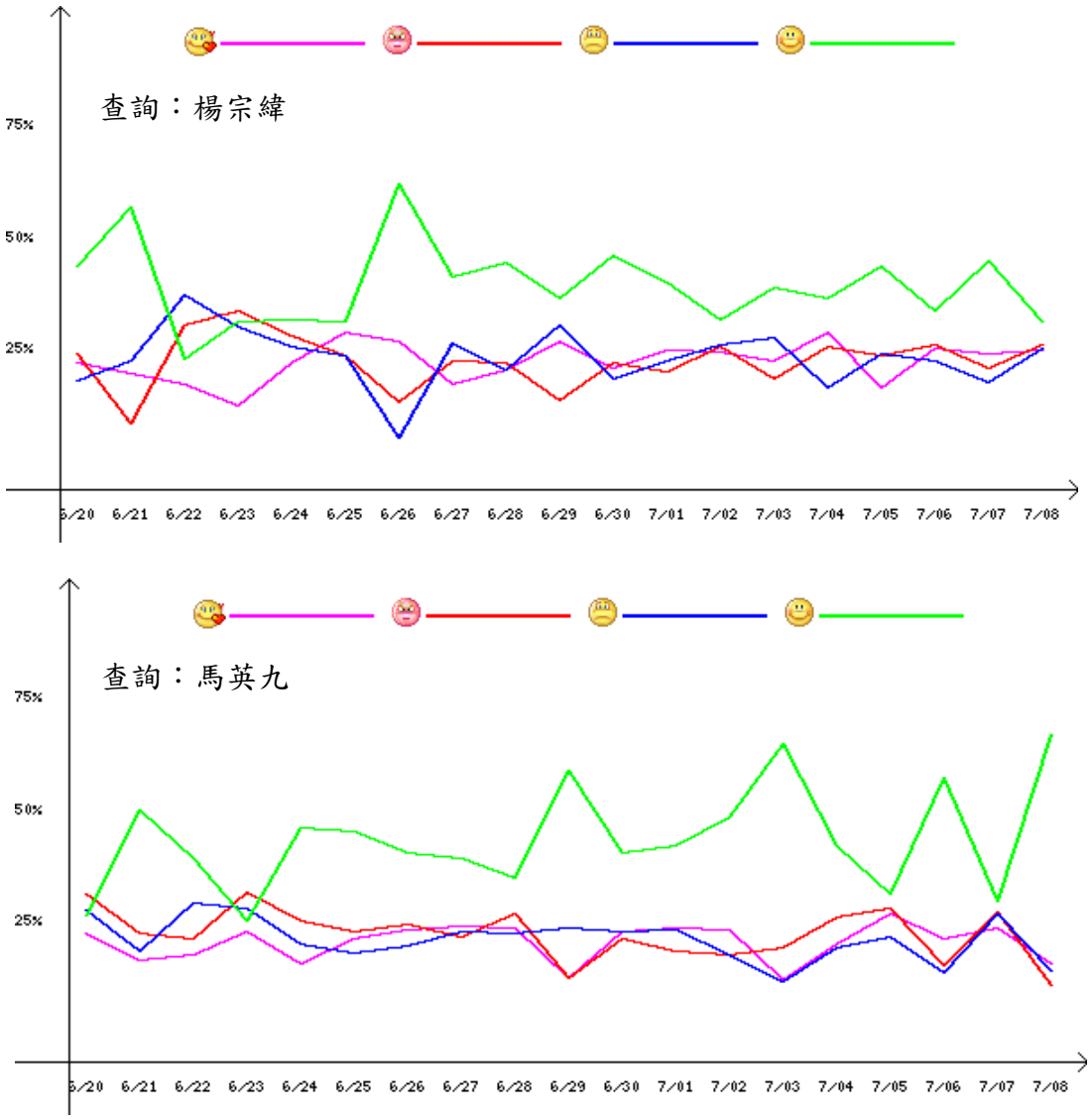


圖四、情緒分析流程圖

[18] http://rpc.weblogs.com/shortChanges.xml

表三、情緒分類器訓練時間分析

訓練句數	訓練倍數	訓練時間	時間倍數
4,000	1	約 6 秒	1
8,000	2	約 25 秒	4
40,000	10	約 31 分	310
120,000	30	約 6 小時 52 分	4,120



圖五、情緒趨勢分析圖

5. 部落格情緒趨勢分析

本文第三節提出一個虛擬部落客「人口普查」概念，可根據部落客名單，以查訪最新文章列表的方式來獲得文本資訊。本節中再提出一個部落客「民意調查」的應用，運

用第四節說明的情緒分析模組，針對個別文本中的各文句進行情緒分析，偵測該文本有無出現喜、怒、哀、樂等情緒，各情緒分數以該情緒的文句數的對數計算之。根據第二節所述，部落格資料庫所提供的文本資訊具有時間標記，因此我們在對文本作情緒分析時，也可以為不同時間域的文本集合計算出對應的情緒分數。

本研究觀察 2007 年 6 月 20 日至 7 月 8 日所發布的文本，個別時間域以天為單位，在對每天的文本集合算出個別情緒的分數比例後，可藉此觀察個別情緒在跨越時間域時的變化走向。圖五顯示以不同關鍵字查詢所獲得的相關文本在跨越時間域時的情緒變化，查詢以不同人物的姓名為例，分別是「楊宗緯」、「馬英九」。圖五所包含的三個子圖，其共通之處在於「樂」情緒類是屬於最大宗的情緒呈現，而「樂」情緒類趨勢的下降通常代表著「哀」情緒類的上升；而「喜、怒、哀」三種情緒類在不同時間域上會互有領先。說明個別的趨勢圖以「查詢：楊宗緯」為例，在 6/20-6/21 呈現較高的快樂的情緒後，6/22-6/23 負面情緒上揚，直到 6/26「樂」情緒類大幅上升後，便持續領先到 7/8 止。

表四、最近時間域情緒分析統計

查詢	喜	怒	哀	樂	說明
楊宗緯	12%	14%	11%	63%	星光幫偶像
林宥嘉	14%	13%	12%	61%	星光幫偶像
股票	7%	11%	9%	72%	台灣股市大漲
貓空	10%	22%	19%	48%	貓纜初期營運負面感受
纜車	11%	21%	18%	50%	貓纜初期營運負面感受
謝長廷	7%	14%	7%	73%	總統參選人
馬英九	14%	15%	10%	62%	總統參選人
王建民	10%	18%	10%	62%	旅美棒球選手
日本	14%	17%	12%	57%	國家、旅遊、演藝娛樂
韓國	17%	21%	15%	48%	國家、旅遊、演藝娛樂

所謂趨勢概念，最令人關心的部分是屬於「目前」、「最近」的趨勢，因此如果設定觀察的時間域為「最近」這個概念時，實作方法可以以時間排序針對各查詢選出若干最新文本，並統計其情緒分數比例。表四即針對各查詢選出最近 150 篇相關文本，在計算文本個別的情緒分數後，列出各查詢綜合相對的情緒比例。

各查詢之間的相對值或極大值，可用來觀察部落客對各項議題不同的感受，以極值的觀察為例，可解讀成部落客最近對股市的表現、或總統參選人謝長廷最感到高興、對貓空纜車的營運最感到不悅與生氣、對韓國則混有嚮往及生氣兩種情緒。以相對值的觀察為例，如楊宗緯和林宥嘉同屬於熱門歌唱比賽出身的偶像，其相關文章較為一致，情緒表現僅有些微的差距。而謝長廷與馬英九雖然同屬於總統參選人，但是其情緒「喜」、「樂」的比例則互有勝負，或許可解讀成儘管部落客普遍對謝長廷的表現較感到滿意，而情感上則較容易喜愛馬英九。

5.1 討論

6. 結論與未來方向

本文探討如何從部落格空間獲得文本資訊，實作出一個資訊系統，並應用在情緒趨勢的分析上。關於資訊系統未來發展方向，一方面可以擴充更多的查訪部落格空間獲得更豐富的文本資訊，一方面可支援如部落格搜尋、部落格摘要等研究議題所需用到的語料。在情緒趨勢的分析上，則可以朝向自動偵測出變動的情緒，並掌握相關的話題的自動機制發展。

參考文獻

Corinna Cortes and V. Vapnik. 1995. “Support-Vector Network,” *Machine Learning*, Vol. 20, pp. 273–297.

Rong-En Fan, Pai-Hsuen Chen and Chih-Jen Lin, “Working Set Selection Using Second Order Information for Training Support Vector Machines,” *Journal of Machine Learning Research*, Vol. 6, pp. 1889–1918, 2005.

Bin He, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang, “Accessing the Deep Web,” *Communications of the ACM*, Vol. 50, 5, pp. 94–101, 2007.

Lun-Wei Ku and Hsin-Hsi Chen, “Mining Opinions from the Web: Beyond Relevance Retrieval.” *Journal of American Society for Information Science and Technology*, Special Issue on Mining Web Resources for Enhancing Information Retrieval, accepted.

Kevin Hsin-Yih Lin, Changhua Yang and Hsin-Hsi Chen, "What Emotions Do News Articles Trigger in Their Readers?" *Proceedings of 30th Annual International ACM SIGIR Conference*, pp. 733–734, Amsterdam, Netherland.

Gilad Mishne. 2005. Experiments with Mood Classification in Blog Posts. *Proceedings of 1st Workshop on Stylistic Analysis of Text for Information Access*.

Gilad Mishne and Maarten de Rijke. 2006. Capturing Global Mood Levels using Blog Posts. *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 145–152.

I. Ounis, Macdonald, M. de Rijke, G. Mishne, and I. Soboroff, "Overview of the TREC 2006 Blog Track," *Proceedings of the 15th Text REtrieval Conference*, Gaithersburg, Maryland, 2006.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen, "Building Emotion Lexicon from Weblog Corpora," *Proceedings of ACL-2007*, poster, Prague, Czech, pp. 133–136, 2007.

Sheng-Chuan Yen, *A Study of Identifying Implicit Trackback in Weblogs and Its Application on Weblog Search*, Master Thesis, National Taiwan University, 2007.

楊昌樺、陳信希。"以部落格文本進行情緒分類之研究，"第十八屆自然語言處理與語音處理研討會論文集，253–269 頁，2006。