台灣大專院校之網路口碑分析-以元智大學為例

謝沛宇
元智大學資訊管理學系
碩士生
s1006205@mail.yzu.edu.tw

張睿晟
元智大學資訊管理學系
碩士生
a0982601000@gmail.com

黃信穎
元智大學資訊管理學系
碩士生
s1006236@mail.yzu.edu.tw

宋瑞蛟
元智大學資訊管理學系
博士生
s979204@mail.yzu.edu.tw

邱昭彰
元智大學資訊管理學系
教授
imchiu@saturn.yzu.edu.tw

摘要

由於近年來少子化問題的影響，大學所面臨的招生競爭越來越激烈，學校風評已成為大學在營運與決策時最重要的參考指標之一。本研究將發展一套自動化意見分析機制，此機制透過擷取網路評論之關鍵字並以啟發式N-phrase 規則將口碑資訊整理與分析。實驗結果顯示在所有議題的評論分類的正確性高達81%以上，以此結果建構定位圖並呈現有價值的資訊。經結果證明，本研究所提之方法可協助各大學了解其於市場定位中之優缺點，進而幫助大學增加市場競爭力。

關鍵詞：啟發式N-phrase規則、意見探勘、口碑分析

Abstract

Because of the trend of fewer children in recent years, universities face the challenge of competitive admissions. The rumor of university has become the most factor in operation and decision-making. This research attempts to develop an automatic web customer opinion analysis system. The system uses crawler to retrieve university-related articles and extract feature words. The heuristic n-phrase rule used to organize and analyze customer generated content information from the Internet. Experimental results show that the heuristic n-phrase rule correctness achieve up to 81% accuracy, which prove that the method could assist universities to understand their advantages and disadvantages in the market position. Finally, we construct positioning map to presents valuable information to assist universities understand their advantages and disadvantages in the market position, thus helping them to increase market competitiveness.

Keywords：Heuristic N-phrase Rule, Opinion Mining, UGC Analysis

1. 前言

由於少子化問題的影響，學校每年畢業生人數逐漸減少，加上我國加入「世界貿易組織 」(WTO) 之後，各級學校招生市場必須開放，其他國家也可到我國招生，各校所面臨的招生競爭越來越激烈，學校如何增加自身的競爭力，獲得學生與家長的認可，進而爭取更多的學生入學就讀，儼然已成為學校經營者目前最重要的課題之一。

在日常生活中，時常會聽到某間學校的風評不錯，因此家長會希望讓小孩到這些風評較佳的學校就讀，而這些風評經由人與人之間口耳相傳後便形成了所謂「口碑」。但是在Web2.0的時代，部落格及社群網站(Social Networking Service, SNS)的興起，訊息傳遞的主控權已快速轉移至一般網友手上。根據尼爾森(Nielsen)發表之「2009年全球網路消費者調查報告」，有七成消費者相信網友在網路上發表的意見與評價，網路口碑往往是消費者尋求產品資訊的優先選擇。而在教育市場也不例外，學生與家長透過網頁搜尋其他學生所提供的學校風評或就讀經驗資訊與討論，以其自身實際經驗、意見與相關知識的分享。

因此學生與家長的評論往往關係著學校的形象，為此學校經營者必須積極且不斷地提升自我服務品質。若能善用這些針對特定主題的網路社群所發表的文章，對學校經營者於其經營將有極大的幫助，同時提高家長與學生滿意程度。然而網路文章數量極為龐大且無固定格式，若要一篇一篇的去閱讀與比較，將耗費大量的人力與時間。因此，本研究目的在發展一套網路意見自動擷取機制，並將口碑資訊整理與分析，以簡單定位圖呈現有價值的資訊。

本研究將於第二節進行文獻探討、第三節討論啟發式N-phrase規則、第四節利用定位圖呈現實驗結果、第五節對本研究進行結論並提出未來發展方向。

2. 文獻探討

現行網路文字探勘中有關意見探勘(Opinion Mining) 的研究大部分仍以英文文章為主體，較少有中文的相關研究。然而在網路使用者中，中文使用者占大多數，因此對於中文的意見探勘研究應更加重視[9]。由於中文與英文在表達方式、語言結構以及詞彙語法上，存在先天上很大的差異，將處理英文的模式直接套用在中文處理並不適宜[5]。以斷詞為例，中文不像英文可以用空白鍵做為斷詞依據，因此，對於中文之斷詞需要另外的工具[2] [4]。

意見探勘可以在特定的議題之下分析出使用者對於該議題的情緒、感受並給予評價，它可以有效地減少人工判讀的工作，並將其轉化為有利於商業智慧或知識管理等用途之上。意見探勘的研究可以將所擷取的評論以Document-based以及Sentence-based來進行分析。由於一篇文章中不可能只討論到一個主題，或是只含一種情緒[4]。因此，這些評論以句子為基礎做分析相較於以文章為基礎做分析會顯的更準確與合理[1][8][6]。

特徵詞擷取是文字探勘中重要的步驟，主要可區分為人工定義與自動擷取兩類。在人工定義方面，例如 Li et al. (2007) 利用已定義好的字典WordNet做為特徵字與褒貶詞的擷取依據[3]；在自動擷取方面，例如 Liu & Hu (2004) 以association rule mining方式找出候選的特徵詞集合，再進行過濾得到最終特徵詞集合[2]。

對於網路使用者者評論分析的主要目的是希望能夠以更簡單、明瞭的方式來替代冗長的文字敘述，讓閱讀者能夠一目了然。Pang et al. (2002) 將關於產品的特徵利用簡短的文字評論以表列的方式呈現[7]；而Liu et al. (2005) 利用柱狀圖顯示每項產品特徵正面與負面傾向的程度，而對於同種產品的比較也可利用此圖清楚呈現[2]。

3. 實驗方法

由於中文語句結構並非如英文使用特定語法及結構所組成，字詞間透過不同的組合便可成為各種意涵的詞，乃屬於發散式的用法並無統一標準。一般使用者於網路文章中所寫作的方式更具有彈性以及自由創作風格，且在網路上不定時便會產生新興文字，在文字處理上亦提升不少難度及挑戰。

為此本研究擬提出一套意見分析機制，透過網路爬蟲程式於網路發問平台上擷取使用者評論元智大學優劣的文章，將其分為學校資源、生活機能、風評、費用、畢業發展以及元智相關等六大議題。經由文章前處理過後再透過本研究所提出的啟發式N-phrase規則找出上述之類別相關用字以及情緒用詞的組合進行分類判定。最後利用定位圖將元智大學在一般大眾心中的評價以視覺化方式呈現，研究流程如圖1，依序分別為：(1)資料擷取與文字前處理 (2)特徵詞擷取及啟發式N-phrase規則 (3)分類評估與視覺化呈現

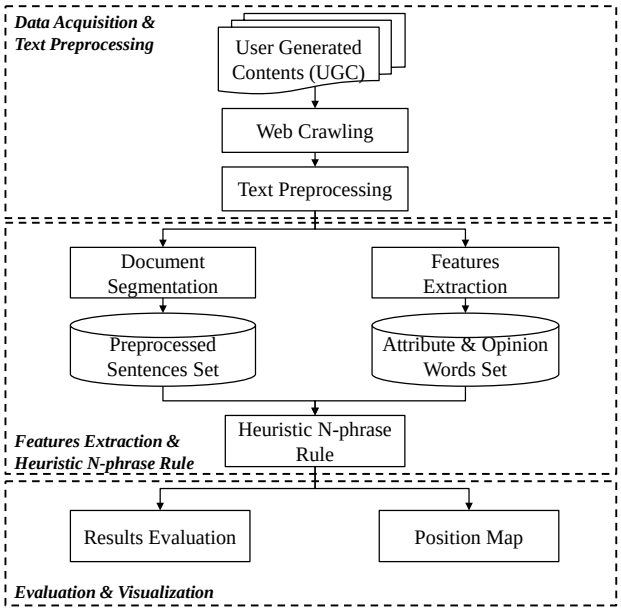

圖 1 研究流程

3.1 資料擷取與文字前處理

3.1.1 網路資料擷取

本研究以元智大學作為研究案例，透過網路爬蟲程式將大多數使用者稱呼元智大學的多種用詞做為關鍵字在奇摩知識家進行搜尋。從搜尋結果中擷取出所需的文章網址、主題、摘要、日期及完整內文等相關資料並存入資料庫。再透過研究人員進行初步過濾將非相關之文件刪除，以得到較具完整性之內容。

3.1.2 文件前處理

將擷取下來的文件透過中研院中文詞庫小組提供的中文斷詞系統(Chinese Knowledge Information Processing, CKIP) 進行斷詞並標註詞性標籤，並藉由中文語法的構造和google suggestion api的搜尋列熱門用詞將CKIP斷詞

結果做比對。將原本斷詞後不具意義的字詞重新合併為富有意涵的詞語。最後，過濾不具資訊性且與研究議題無關的停用字(Stop Word)，如代名詞、介係詞、嘆詞等，而保留較重要的詞性如名詞、動詞、形容詞等。此外由於中文字詞在使用上較為紊亂無一致性，在處理時常會因為數量過多而造成運算以及判讀上的效率不彰。Document Frequency (DF) 則是計算一個字詞所出現的文章篇數，經常被用於字詞的前置篩選，以便減少多餘的雜訊並且降低字詞數量。在本研究中則保留DF>=2的字詞以進行後續處理。

3.2 特徵詞擷取與啟發式 N-phrase 規則

3.2.1 文件分句

由於網路使用者在奇摩知識家詢問大學評價時，回覆者的內容通常是以多面向的觀點進行分析，因此涉及多議題的內容。若以整體內容作為基礎，將該內容歸類為某一特定議題，則容易失去資料之真實性，且不利於分類判定。為此本研究將常見的標點符號如:逗號、句號、分號、驚嘆號及問號等取代為自行定義之標籤，另外由於網路文章常使用空白以及換行字元作為語句的結束，因此亦列為取代項目。透過自行定義標籤將文章做分割之程序，並依照每句中所談論到的議題及評價標記並歸類。

3.2.2 特徵詞篩選

將經由文件前處理過後的相關文章與字詞進行Term Frequency (TF)計算，進而得到字詞於文件中出現的次數。經過初步過濾後，就其結果挑選出與六大議題相關之字詞製成屬性詞詞集和情緒詞詞集。此外本研究亦採用台灣大學自然語言處理實驗室所建立的語意辭典NTUSD作為情緒詞詞集，以提升本研究情緒用詞之完整性。

由於中文文字的多樣性，字與字的組成方式不同，形成類似或是完全不同的語意。因此我們利用已定義好的屬性及情緒詞集與文件分句後的句子進行比對，按照議題內容將句子中出現的特徵詞以六大議題名稱(學校資源、生活機能、風評、費用、畢業發展以及元智相關)取代；情緒面意見詞語(正面、負面)的處理方式亦是如此，透過特徵詞取代可使後續處理時能增加效率且清楚明瞭。如:師資頂尖 → 【學校資源】【正面】、生活圈荒涼 → 【生活機能】【負面】、學費貴 → 【費用】【貴】，相關取代用詞的分類表如表1。

表1 各議題及情緒用詞分類表

議題	特徵詞
學校資源	校地、校舍、獎學金、師資、硬體、設備…
生活機能	生活圈、交通、住宿、飲食、宵夜、夜市…
風評	評鑑、評價、評比、排名、風評、名聲…
費用	學費、學分費、雜費、住宿費、開銷…
畢業發展	畢業生、企業界、起薪、主管、出路…
元智相關	元智、校園、辦學、私立、學生…

情緒面	特徵詞
正面	優良、頂尖、齊全、推薦、完善、讚許…
負面	荒涼、墮落、蕭條、老舊、狹小、差、貴…

此外否定字對於語句會造成完全相反的結果，因此本研究亦針對否定字進行處理。常見的否定詞為【不】和【沒有】，當情緒詞前面一個字詞為否定詞時，則將該字詞的正負面情緒進行轉換。如：設備不齊全 → 【學校資源】【不】【正面】→ 【學校資源】【負面】、學費不貴 → 【費用】【不】【負面】→【費用】【正面】。

3.2.3 啟發式N-phrase規則

本研究所提出的Heuristic N-phrase規則，以Sentence-based的觀念將經過前置處理的詞句中六大議題字與情緒面用字間的距離減少，進而將其合併。啟發式N-phrase規則如下：以詞句中出現的議題字為中心點，檢驗其前後N個字詞內是否包含情緒面用詞，若於N個字詞內找到情緒面用詞，則將議題字和情緒面用詞連接並合併。由於中文用法中形容詞較常出現於名詞之後，因此啟發式N-phrase規則的檢驗順序為先後再前，依此類推找出N個範圍內的議題和情緒詞組合。透過啟發式N-phrase規則的方式可以藉由更接近人的寫作模式找出每個語句中所表達的議題以及情緒為何，其相關的範例與虛擬碼如圖2及圖3。

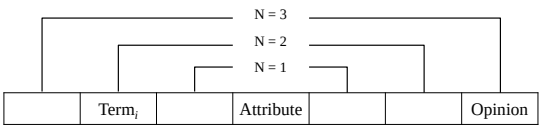

圖2 啟發式N-phrase規則

Input: sentence_set, attribute_set, and opinion_set
Output: rule_list
Procedure: Heuristic n-phrase rule
1. Pattern with length n
2. Extract the phrases either side of "attribute"
3. For each sentences s_j in sentence_set do
4. For each phrases p_i in s_j do
5. If phrase p_i exists in attribute_set then
6. For x=1 to n-1 do /* Once an opinion word is found, exit for loop*/
7. 1. Examine if the phrase p_{i+x} is an opinion word;
8. 2. Examine if the phrase p_{i-x} is an opinion word.
9. End for
10. End if
11. Insert the results of heuristic n-phrase rule into rule_list
12. End for
13. End for

圖3 啟發式N-phrase規則之虛擬碼

3.3 評估與視覺化呈現

3.3.1 分類評估

經由本研究所提出的啟發式N-phrase規則找出語句中的議題與情緒詞組合後，計算每個句子和各特徵詞組合的TF-IDF值，並將其關係以向量空間模型表示。以C4.5、Naïve Bayes (NB)、Support Vector Machine (SVM) 等常用的分類工具計算其分類的正確性藉以驗證模型的實用性，並依照分類結果進行深入探討。分類正確率計算方式如下式：

$$\text{正確率} = \frac{\text{各類別實際與預測相符之語句總數}}{\text{全部語句總數}}$$

3.3.2 視覺化呈現

經過啟發式N-phrase規則處理過後的句子，整理網路使用者對於元智大學在學校資源、生活機能、風評、費用、畢業發展以及元智相關的正負面評價。透過對應分析(Correspondence Analysis, CA) 製作定位圖，將多維度表示的列聯表呈現於二維的平面上。能以簡單明瞭的定位圖了解使用者對於元智大學各項議題的看法，針對評價差的議題進行改進。

4. 實驗與結果

4.1 資料描述

本研究資料來源為奇摩知識家，並以「元智」相關用詞作為搜尋研究相關資料之關鍵字。藉由其關鍵字搜尋之文章總數為907筆，將其分割為9352個詞句。經由人工標記後，同時具備議題分類及正負面評價的句數共有478句，其中正面評價佔426句，負面評價佔52句，資料日期區間為西元2004年12月20日至2012年11月30日。

4.2 分類評估

本研究以C4.5, Support vector machine (SVM), Naïve Bayes (NB) 常用分類工具計算啟發式N-phrase規則的分類正確率，分別檢驗N=2,3,4等結果。在所有議題評論的褒貶分類結果，正確性有相當好的成果，10-fold 平均分類正確性高達 81%以上。見表2。

表2 N-phrase分類正確率

	Bi-phrase (N = 2)	Tri-phrase (N = 3)	Four-phrase (N = 4)
	Accuracy Results		
C4.5	83.18%	84.76%	85.58%
SVM	83.32%	84.84%	85.84%
NB	81.61%	83.07%	84.07%

* Testing results based on 10-fold average

4.3 市場定位圖

在討論元智大學的口碑議題上，將網路對於元智的評價歸納為六項議題，分別是「學校資源」、「畢業發展」、「生活機能」、「費用」、「風評」與「元智相關」等六個項目，透過列聯表可以從中了解網路評論對於其六項議題的正負面評價，列聯表呈現如下表3所示。

表3 各項議題之列聯表

議題 / 情緒	學校資源	畢業發展	生活機能	費用	風評	元智相關
正面	46	26	9	3	47	163
負面	4	1	3	10	2	20

同時，利用CA製作市場定位圖如圖4，而其中了解元智大學在六項議題中，有四項都是偏向正面評價，分別為「學校資源」、「畢業發展」、「風評」、「元智相關」，說明了元智大學無論在提供學生的使用資源上、教授的教學品質、畢業之後的發展、品牌形象等等皆是獲得好評；而「生活機能」項目較偏向於中立，可能是由於元智大學鄰近的食衣住行等生活機能相較於其他學校，無法使人有深刻的感受及體驗；唯有在「費用」項目評價較不盡理想，從奇摩知識家中擷取與「元智」關鍵字相

關的資料中可以得知，導致「費用」項目較偏向負面的原因，主要述及的部分是學費稍貴，與他校比較下來就讀的費用差異稍大，不僅隸屬私立的元智大學的學費高於各國立大學，元智大學與其他私立大學學費對比也較為昂貴，新生在選擇入學的學校時，就讀期間所需的費用也是考量之一。

因此元智大學若是欲提高其招生率及提升競爭優勢，建議以「費用」項目為優先考量，其次從「生活機能」項目著手，希望能夠有所助益。

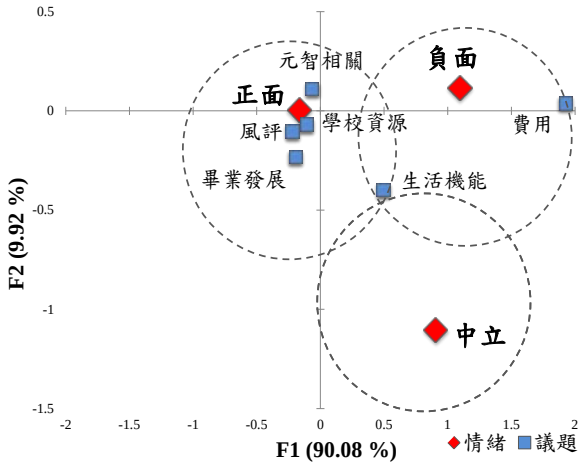

圖4 元智大學市場定位圖

5. 討論與結論

網路中對於元智大學的評論，不論是對學校自身或是欲了解學校口碑的網路使用者，皆是具有參考的價值。由元智大學市場定位圖4可以得知，六項議題項目中，「學校資源」、「畢業發展」、「風評」、「元智相關」四個項目在網路的評價上皆是偏於正面；而「生活機能」項目較偏向於中立；而偏向負面評價的項目，只有「費用」項目稍不夠理想，而主要述及的原因是學費支出較高，因此元智大學欲增加競爭優勢，建議以「費用」項目作為優先討論項目，其次再從「生活機能」項目著手，希望此實驗結果對於元智大學能夠給予實質幫助。

本研究發展啟發式N-phrase規則，將網路上針對元智大學評價，透過完善的流程處理在中文環境下文件中包含多議題之意見意向所面臨到的問題。同時以常見的分類方法衡量本規則之分類正確性，結果顯示本研究提出的啟發式N-phrase規則同時具備高效率及準確性的結果。

未來研究將蒐集其他大專院校之網路口碑，透過啟發式N-phrase規則將其進行意見分析，並以CA圖進行呈現。除了讓各大專院校瞭解自己的優劣勢，並可比較同地區或者是排名相近的其他大專院校，透過分析結果則可提出改進方向，藉此領先其他競爭者。

此外將對本研究預先定義的議題發展出自動化特徵詞選擇模式來取代現行人工篩選模式。此外，目前啟發式N-phrase規則僅處理單純正負面情緒，未來將對褒貶義詞的褒貶強弱程度給予不同級別，以區分情緒的差別。

參考文獻

[1] Hu, M., Liu, B., "Mining and summarizing customer reviews," ***Proceeding of Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining***, pp. 168-177, 2004.

[2] Liu, B., Hu, M., Cheng, J., "Opinion Observer-Analyzing and comparing opinions on the web," ***Proceeding of the 14th international Conference on World Wide Web***, pp. 342-351, 2005.

[3] Li, J., Sun, M., "Experimental study on sentiment classification of Chinese review using machine learning techniques," ***Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*** , No. 4368061, pp. 393-400, 2007.

[4] Li, Z., Zhang, M., Ma, S., Zhou, B., Sun, Y., "Automatic extraction for product feature words from comments on the web," ***Lecture Notes in Computer Science***, Vol. 5839, pp. 112-123, 2009.

[5] Li, S., Ye, Q., Li, Y.J., Law, R., "Mining features of products from Chinese customer online reviews," ***Proceeding of Journal of Management Sciences in China***, Vol. 12, No. 2, pp. 142-152, 2009.

[6] Li, L. , Yao, T., "Kernal-based sentiment classification for Chinese sentence," ***Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology***, pp. 27-32, 2007.

[7] Pang, B., Lee, L., Shivakumar Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques," ***Proceedings of the ACL-02 conference on Empirical methods in natural language processing***, pp. 79-86, 2002.

[8] Popescu A-M, Etzioni O., "Extracting

product features and opinions from reviews," ***Proceeding of HLT-EMNLP***, pp. 339-346, 2005.

[9] Zheng, W., Ye, Q., "Sentiment classification of Chinese traveler review by support vector machine algorithm," ***Proceedings of the 3rd international conference on Intelligent information technology application***, pp. 335-338., 2009.

[10] CKIP (Chinese Knowledge and Information Processing)：
http://ckip.iis.sinica.edu.tw/CKIP/, 1986.