發展植基於支撐向量機的語意導向指標以分類部落客之情緒

陳隆昇
朝陽科技大學
資訊管理系
助理教授
lschen@cyut.edu.tw

邱慧如
朝陽科技大學
資訊管理系
研究生
s9614621@cyut.edu.tw

摘要

對於一些以文字內容為基礎的溝通工具(如部落格)，情緒的辨識十分重要。在商業部落格中，部落客所給予產品的負面內容或評價將會在網路空間非常迅速地散播。此外，在部落格圈(Blogosphere)中揭露一些不為人知的企業內幕，將會嚴重影響到公司的商譽。這些負面的訊息通常會給傷害企業帶來重大的損害。近年來，許多學者專注於語意分類(Sentiment Classification)的研究上，希望可以有效地辨識客戶的負面情緒，以協助企業回應。因此，本文結合機器學習技術和資訊檢索之優點，提出一個植基於支援向量機的語意導向指標(SVM based SO index)，以協助公司迅速地、有效率地偵測具有傷害的負面部落格內容。實驗結果顯示，所提的方法優於類神經網路為主的語意導向指標及數種傳統語意導向指標。

關鍵詞：情緒導向、支援向量機、情感分類、情緒運算、部落格

Abstract

Recognizing emotion is extremely important for some text-based communication tools such as blogs. On commercial blogs, bloggers' negative comments or evaluations of products spread quickly in the cyber space. Moreover, an exposure of an inside story in the blogosphere may influence a company's reputation. These negative comments are often harmful to enterprises and might result in great damage. Recently, researchers have paid much attention on sentiment classification to efficiently identify customers' negative emotions for helping companies to carefully response customers' comments. Following this trend, this study proposed a support vector machines based semantic orientation index (SVM based SO index) which integrates the advantages of machine learning techniques and Information Retrieval (semantic orientation indexes) to help companies detecting harmfully negative bloggers' comments quickly and effectively. Experimental results indicated that our proposed SVM based SO index outperform traditional approaches, NN based SO index and several semantic orientation index.

Keywords: **Semantic Orientation, Support Vector Machines, Sentiment Classification, Affective Computing, Blogs**

1. 緒論

90 年代，部落格迅速地成為受歡迎的網路溝通工具(Chen et al., 2008)。依研究顯示，目前已經有 60 億萬個部落格被建立(Singh et al., 2008 ; Murphy, 200)，且以每天新建 50,000~70,000 個部落格的速度持續增長(Singh et al., 2008；Martin, 2005)。部落格是一種運作於網路或網際網路的伺服器上的(server)軟體程式；使用者可以透過任何電腦存取部落格的內容。部落格和網頁很相似，不過差別在於部落格具有讓使用者自行撰寫、更新網頁內容之特性。部落格，曾被定義成"cross between a diary, a web site, and an online community"；也因此部落格具備了公開分享資訊或意見的特性(Yap et al., 2005)。

在商業領域中，最初，組織是利用部落格做為專案管理的工具，並且用來協同組織中各個功能領域。但因部落格具有公開分享資訊或意見之特性，也因此改變了企業以及企業和客戶之間的關係(Singh et al., 2008)。瞭解部落客的情緒，不但可以做為企業、組織做為新穎的行銷工具，針對部落格撰寫之內容做最佳的回應(Wu et al., 2006, Subasic et al., 2001)，並且可以用來瞭解競爭對手之行銷概況和即時地過濾對於企業形象會造成傷害之負面評論。所以

瞭解部落客撰寫內容的情緒探勘議題，在近幾年也成為焦點(Gilly et al., 2006)。

部落格圈 (blogosphere)，包含部落格(weblogos)和部落格相關網頁，對於行銷人員、社會心理學學者和其它在擷取和探勘意見、評論、情緒和屬性的研究，部落格具有豐富的資源(Mishne et.al, 2006)。企業可以藉由部落格內容，瞭解企業的未來趨勢。而且在企業部落格中，一個熱門的主題，它可能會影響到產品的生命週期(Chen et al., 2008)；而且近幾年，利用部落格做為企業行銷工具的風氣也愈來愈盛行。例如：預算不高的小成本電影-海角七號，藉由部落格行銷方式，讓網友透過網路文字相傳，也締造了亮麗的票房成績。相同地「水能載舟，亦能覆舟」，負面的評價，透過部落格快速且大量地傳播，亦容易對企業造成極大且難以撫平之傷害。例如：eBay 線上購物網站，曾經因為賣家出貨速度太慢，導致賣家透過撰寫評價抨擊買家，這對買家而言也是種莫大的傷害。而且在企業部落格中，揭露一些不為人知的內部秘辛或對企業產品做出極為負面之評價，都對會企業的商譽造成影響(Chen et al, 2008)。因此，我們選擇部落格做為本研究之對象。除了協助企業過濾負面影響之內容，把對於企業有害之負面資訊做過濾。並且藉由部落格情緒探勘，瞭解部落客對於企業、產品或服務之評價。提供潛在客戶對於企業或產品、服務有更進一步的認識。並且協助行銷人員，依據客戶撰寫之內容，瞭解公司產品、服務之滿意度。

自動化辨識人類情緒為情緒運算的目標，辨識處理不但富於我們對於辨識的瞭解，也協助我們對於人類的輸入去做反應(Leshed et al., 2006)。常見的情緒辨識方法。大致上分成非監督式學習法(語意導向指標)與監督式機器學習法(Ye et al., 2006；Turney et al., 2003；Turnet, 2002；Chaovalit et al., 2005；Fei et al., 2004)。非監督式學習法具有即時使用之優點，但計算之精確率通常較監督學習法低(Chaovalit et al., 2005)。而監督式機器學習法具有較高的判定正確率，但訓練十分耗時。因此，本研究將結合兩者之優點，提出一個基於支撐向量機(Support Vector Machines, SVM)的語意導向指標(SVM based SO index)，希望藉此可以協助企業迅速地、有效率地偵測具有傷害的負面部落格內容，以降低負面內容及評論對於企業的所給予傷害。

2. 文獻探討

過去相關研究指出，自動化情緒分類可以將不同的客戶產品評論分成正面/反面(Turney et al., 2003；Turnet, 2002；Chaovalit et al., 2005；Fei et al., 2004)，用以協助潛在客戶對於產品之瞭解。在搜尋引擎中，也允許使用者針對需求之評論做詳細的主題或正/反導向之詳細說明(文章為推薦或不推薦以及文章為正面或反面)，使瀏覽使用者查詢資料更加便利與迅速(Turnet, 2002)。另外 Tong(2001)的 *sentiment timelines* 系統，可以用來追蹤電影與角色伴演的相關線上討論之評論(Tong, 2001) 。

在自動化情緒分類中，常見的分類方法。大致上分成非監督學習法(例如：語意導向指標-PMI 等)與監督機器學習法(例如：n-gram classifiers) (Ye et al., 2006；Turney et al., 2003；Turnet, 2002；Chaovalit et al., 2005；Fei et al., 2004)。非監督學習法的優點為適合用於即時使用之應用，而缺點則是計算之結果精確率通常較監督學習法低(Chaovalit et al., 2005)。過去研究顯示，依據主題為主的內容分類研究中，監督機器學習分類法是種常見的技術，而且分類結果也都有不錯的成效。但是將監督機器學習分類法應用於情感內容為主的分類研究中，分類的成效比以主題為主的內容分類研究差(Pang et al., 2002) ；而且監督機器學習之方法，訓練過程中必需花費許多時間訓練分類模型，因此較費時(Chaovalit et al., 2005)。所以，本研究將結合非監督學習法與監督機器學習法之優點，來改善傳統語意導向之分類結果。

2.1 非監督式學習法

2.1.1 關連式語意導向

『關連式語意導向』(Semantic Orientation from Association, SO-A)為一種常見的語意導向推論概念。隨後所介紹的『語意導向 PMI』及『語意導向 LSA』，皆由 SO-A 的概念做延伸的推論。SO-A 主要是由利用語意的關聯計算語意導向。在 SO-A 中，需先建立一個正向與負向的模型集合(paradigm set)。隨後計算文字或片語與正向模型集合中的文字或片語之關聯，減掉文字或片語與負向模型集合中的文字或片語之關聯。若加總結果為正，表示此兩個文字或片語有正向的語意導向；若結果為負，則表示兩個文字或片語有負向的語意導

向。SO-A 公式如下：

$$\text{Pwords}= \text{a set of words with positive semantic orientation} \quad (1)$$

$$\text{Nwords}= \text{a set of words with negative semantic orientation} \quad (2)$$

$$A(word_1, word_2)= \text{a measure of association between } word_1 \text{ and } word_2 \quad (3)$$

$$SO-A(word) = \sum_{pword \in Pwords} A(word, Pword) - \sum_{nword = Nwords} A(word, Nword) \quad (4)$$

公式(1)乃為正向模型集合，公式(2)為負向模型集合。在本計劃書中的範例集合分別為：Pwords={feel good, good movie, good performance}、Nwords={film noir, long time, old fashioned}。公式(3)是藉由公式(4)計算字與字或片語與片語之間的關聯。當 A(word$_1$,word$_2$)計算結果大於零，表示此文字具有正向的語意導向；若計算結果小於零，則表示此文字具有負向的語意導向。

2.1.2 語意導向 PMI

『語意導向PMI』演算法(Semantic Orientation from PMI, SO-PMI)是一種利用Pointwise Mutual Information計算文字或片語之間的語意導向。主要是透過資訊檢索(information retrieval；IR)hit的數量，以獲取文字與文字之間一起發生的頻率(Turney et al.,2003)。隨後，再藉由SO-A之概念做延伸的計算，以求出語意導向。SO-PMI經常被用於計算語意導向之相關研究中(Turnet, 2002；Turney et al., 2003；Chaovalit et al., 2005；Abbasi et al., 2007)。PMI方法中，word$_1$與word$_2$之間計算公式，定義如下：

$$\text{PMI}(\text{word}_1, \text{word}_2)=\log_2\left(\frac{\text{P}(\text{word}_1 \text{ Near word}_2)}{\text{P}(\text{word}_1)\,(\text{word}_2)}\right) \quad (5)$$

P(*word*$_1$ Near *word*$_2$)表示word$_1$與word$_2$一起發生的機率。假設word是統計獨立(statistically independent)，則透過P(*word*$_1$) P(*word*$_2$)相乘取得一起發生的機率。隨後P(*word*$_1$ Near *word*$_2$)與P(*word*$_1$) P(*word*$_2$)的比例，即可做為衡量word之間的關聯程度。最後公式(5)所計算之數值，若加總結果為大於零表示此兩個文字或片語有正向的語意導向；若結果為小於零則表示兩個文字或片語有負向的語意導向，公式如下：

$$SO-PMI(word) = \sum_{pword \in Pwords} PMI(word, pword) - \sum_{nword \in Nwords} PMI(word, nword) \quad (6)$$

PMI-IR 計算 PMI 是透過搜尋引擎(因此也稱為 PMI-IR)hit 的數量(matching documents)。在許多 SO-PMI 相關的研究中，經常使用 AltaVista(Turnet, 2002；Turney et al., 2003；Read, 2004)、Yahoo(Mishne, 2005)等搜尋引擎做為研究工具。而在本研究中，我們選擇了 AltaVista 做為檢索的研究工具。主要除了 AltaVista Advanced Search engine5，大約有三十五億個網頁量；另外 AltaVista Search engine5 也具備了 NEAR 運算元之功能。NEAR 運算元可以使搜尋引擎在 2 個字彙之間，相間 10 個字彙之內容做檢索。而且在先前的相關研究中，也顯示了透過 NEAR 運算元之計算結果較 AND 運算元佳(Turney et al.,2003)。

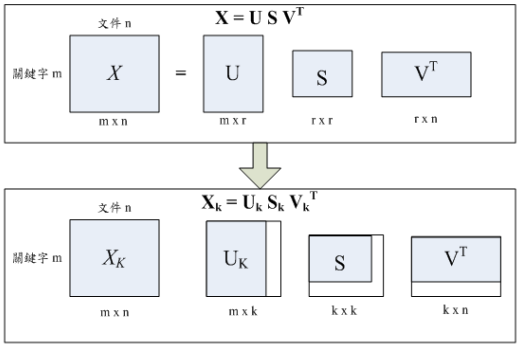
2.2.3 語意導向 LSA

『語意導向LSA』(Semantic Orientation from LSA, SO-LSA)是利用潛在語意分析(Latent Semantic Analysis)計算文字之間的語意關聯強度(Landauer and Dumais, 1997)。LSA透過奇異值分解(singular value decomposition；SVD)，將文集中的文字做分析與統計之間的關係。首先，利用文章內容建構一個詞彙-文件矩陣(Term-Document Matrix；TDM) *A*。矩陣中橫向量表示關鍵字，直向量表示文件。而在此矩陣中的每一個細胞(cell)，表示一內容中關鍵字之權重。通常在SO-LSA中會透過典型的TF-IDF詞彙加權技術，計算細胞中的權重。

隨後透過奇異值分解，將 *A* 矩陣分解成三個不同的矩陣 USV^T，如圖一所示。矩陣 $\boldsymbol{S} = \boldsymbol{diag(\sigma_1,\ldots,\sigma_p)}$且 $\boldsymbol{p = min\{m, n\}}$為含有奇異值的對角矩陣(diagonal matrix)。上述之奇異值可被表示為 $\sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_r \geqq \sigma_{r+1} \geqq \ldots \geqq \sigma_p \geqq 0$，當中的 *r* 為矩陣 *A* 的秩(rank)；而矩陣 **U** 與矩陣 **V**，為 *m*×*m* 及 *n*×*n* 的正交矩陣 (Golub and Van Loan, 1989)。若矩陣 ***A*** 的秩(rank)為 ***r***，則對角矩陣 ***S*** 會有 ***r*** 個不為 0 的值。矩陣 U 與矩陣 V 是由 AA^T 與 A^TA 的前 r 個非零特徵值所對應特徵向量組合而成。假設 $\boldsymbol{S_K}$ 且 $\boldsymbol{k<r}$，為前 *k* 個奇異值所組成的對角矩陣，而 U_k 和 V_k 則為 U 和 V 的前 *k* 個行向量。矩陣 $U_kS_kV_k^T$ 為透過 SVD 而縮減維度的矩陣。U_k 為維度縮減後的詞彙向量，相似的兩個詞彙可以藉由兩詞彙向量夾角

大小的餘弦值(cosin value)來衡量。隨後利用公式(7)將兩個詞彙做計算，若加總結果為大於零表示此兩個文字或片語有正向的語意導向；若結果為小於零則表示兩個文字或片語有負向的語意導向，公式如下：

$$SO-LSI(word)=\sum_{pword\in Pwords} LSA(word,pword)-\sum_{nword\in Nwords} LSA(word,nword) \quad (7)$$



圖一、奇異值分解(Deerwester et al., 1990)

2.2 支撐向量機

在機器學習方法中，Joachims(1998)和 Yang et al.,(1999)曾經利用支援向量機(Support Vector Machines; SVM)於主題內容分類的實驗中，而且精確度皆達到 85%以上。因此，本研究採用支撐向量機為本研究之學習方法。

2.2.1 線性支撐性量機

在 Burges (1998)等學者的研究中，曾經說明線性支援向量機(Linear Support Vector Machines)如何區分二類別的資料。線性支援向量機是先對於每筆不同的訓練資料做標註"+1"或是"-1"，以數學公式表示如下：

$$(x_1,y_1),\ldots,(x_i,y_i),\quad x_i\in R^d,\ y_i\in\{-1,1\} \quad (8)$$

若有一超平面可將標註"+1"或是"-1"的二類別資料做區分，則此超平面稱為區分平面(Separating Hyperplane)。隨後定義區分平面的邊界(margin)為 d_++d_-、d_+ 與 d_-，分別為標註"+1"與標註"-1"兩類別的訓練資料與區分平面的最短距離。線性支援向量在處理區分為二類別資料時，會尋找最大邊界的區分平面。而此類型的資料必須符合下面的限制條件公式：

$$x_i.w+b\geq+1 \quad for\ y_i=+1 \quad (9)$$

$$x_i.w+b\leq-1 \quad for\ y_i=-1 \quad (10)$$

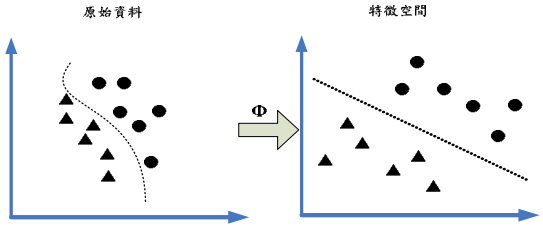
公式(9)(10)可結合為以下之不等式：

$$y_i(x_i.w+b)-1\geq 0\ \forall i \quad (11)$$

由限制公式(9)(10)可獲得 $d_+=d_-=\frac{1}{\|w\|}$，故邊界為 $\frac{2}{\|w\|}$。假設要尋找最大邊界的區分平面，就需在符合限制公式(11)的條件之下，求出最小值。若限制公式(11)中等號成立，則稱為支援向量機。

2.2.2 非線性支撐性量機

線性支援向量機(Linear Support Vector Machines)是藉由找出線性區分超平面，來進行分類之研究。但並非所有的資料皆可以找出線性區分超平面。因此針對於非線性的問題，即可透過非線性支援向量機來解決。非線性支援向量機是將原始資料藉由映射函數 Φ 映射(mapping)至高維度的特徵空間(feature space)裡，隨後再進行線性的分類。圖二為原始資料映射到特徵空間的示意圖。



圖二、原始資料映射到特徵空間示意圖

2.2.3 核函數

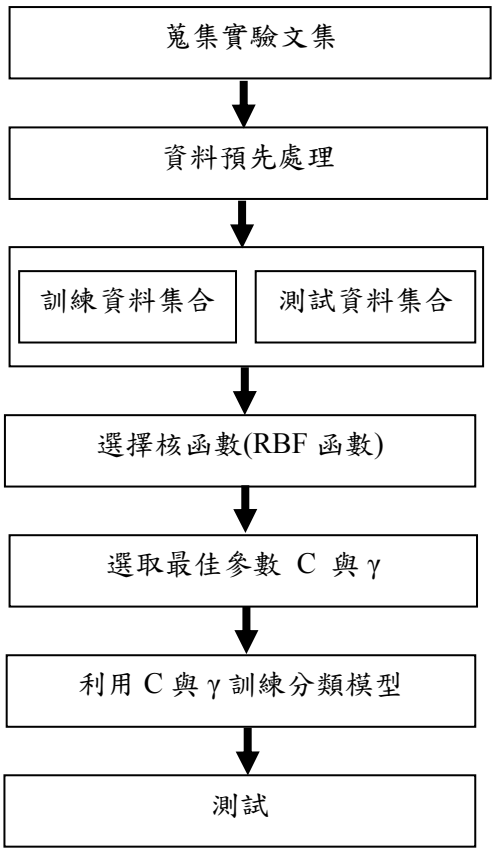
在非線性的問題中，SVM 將原始資料的輸入向量映射到高維度的特徵空間裡，隨後在新的特徵空間中將類別資料做分區分。所以我們可以不用知道如何映射，只需要選擇合適的核函數(Kernel)。核函數主要是用來處理兩個向量在特徵空間中的內積值(inner product)。常見的核函數如表一所示：

表一、常見核函數種類

名稱	公式
線性函數(Linear)	$K(x_i, x_j) = x_i * x_j^T$
多項式函數 (Polynomial kernel)	$K(x_i, x_j) = (1 + x_i * x_j)^d$
RBF函數 (Radial basis function kernel；RBF)	$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$
S型核函數 (Sigmoid kernel)	$K(x_i, x_j) = \tanh(kx_i * x_j - \delta)$

註：γ ,r, *d*為參數，皆由使用者自訂及調整(Hsu et al., 2003; Gumn, 1998)

每個核函數皆有不同的參數，必需由使用者依據不同的核心函數做設定及調整。參數和高維度特徵空間之結構與最後解的複雜度有高度的關係，故核函數的參數設定必須依據不同的問題，選擇合適的核心函數。上述所列出之核函數中，以放射型函數為最常見。過去Smola(1998)的研究中指出，若對於資料的性質沒有做事先的瞭解時，使用 RBF 型函教可以獲得較佳的預測結果，因此本研究亦是利用 RBF 型的函數於實驗中。



圖三、基於 SVM 之語意導向指標實驗流程

3.植基於支撐向量機之語意導向指標

在此章節，我們將說明本文所提出之基於 SVM 為主的語意導向指標，將部落格之評論做情緒分類的實驗流程。圖三為實驗流程，共有 7 個步驟，敘述如下：

步驟1：蒐集文集

透過網際網路(Internet)於各大部落格或評論網站，蒐集各種不同領域之文字為主的部落格或評論內容或文章。

步驟2：資料預先處理

Step2.1：藉過資料預先處理之步驟，將文集轉換成詞彙-文件矩陣。並且，透過第二章計算各種語意導向方法之語意。

Step2.2：建構欲輸入至支援向量機之指標矩陣。

步驟3：轉換資料格式及訓練集合和測試集合

Step3.1：計算四種傳統語意導向之數值，建構成輸入至支援向量機中的格式。另外，本研究也會將語意導向數值，藉由-1、0和1的處理，形成基於SVM的正規化語意導向指標及使用詞彙向量矩陣之分類結果，與本研究所提出之方法做比較。

Step3.2：將所有欲輸入(input)至支援向量機的語意指標矩陣，分成訓練集合(training sets)和測試集合(testing sets)，並且轉換成支援向量機所需之資料格式。

步驟 4：選擇核函數

Step4.1：本研究使用林智仁博士所開發的 LIBSVM2.86 做為分類之工具，而核函數預設值為 RBF。

步驟 5：選取最佳 C 與γ參數

Step5.1：利用 grid.py 選取最佳 C 與γ的參數。

步驟 6：訓練分類模型

Step6.1：利用步驟 5 所選取之最佳的 C 與γ參數，訓練分類模型。

步驟 7：測試

Step7.1：測試分類模型之效率。

4. 實驗

4.1使用文集與資料前處理

本實驗目前以電影評論[1]為研究文集。此資料庫共有 2000 筆文字檔案，1000 筆為正面情緒之評論；另 1000 筆為負面情緒之評論。過去許多內容探勘、情緒探勘之研究中(Turnet, 2002；Chaovalit et al., 2005；Zhuang et al., 2006)，經常利用電影評論來做為實驗文集。

資料前處理(Data Preprocessing)此步驟中，我們需利用文章中出現較多次的關鍵字來建構一個詞彙-文件矩陣。本實驗中，利用 Rubryx2.0[2] 來統計關鍵字的出現頻率，。Rubryx2.0 是一套共享軟體，它是一套以 N-Gram 為基礎，所構成的分類器。因為文章中的所有文字，並非全部皆可以使用於語意分類的研究中。過去 Hu 與 Liu(2004)曾指出，客戶評論中的名詞和名詞片語可以視為一個特徵(feature)，而當中的形容詞通常可以用來表達意見和情感。而且 V. Hatzivassiloglou 和 J.M. Wiebe (2000) 認為形容詞是主觀的，在評估句子時也是一種良好的特徵。但是 Turney(2002)指出，雖然單一的形容詞可以用來做為主觀的指示，但是它乃不足以決定語意導向。例如：形容詞“unpredictable”在汽車領域中，具有反面的語意，但是“unpredictable steering”在電影領域中，卻具有正面的語意。因此擷取特徵字時，最好是擷取兩個連續的文字。POS(Part of Speech)標籤，是一種擷取連續兩個字的規則。Brill (1994) 是第一個將 POS 標籤應用於評論實驗中的學者。而 POS 標籤如表二：

表二、Part-of-Speech 之規則

	First word	Second word
a.	Adjective	Noun
b.	Adverb	Adjective
c.	Adjective	Adjective
d.	Noun	Adjective
e.	Adverb	Verb

過去 Turney(2002), Wang et al(2005) 及 Chaovalit & Zhou(2005)的研究中，皆使用POS標籤，做為指示語意導向的特徵依據。故本研究，也利用POS標籤來做為擷取關鍵字之規則。因為文章中會出現一些常出現的Stop Words，也因此我們將移除Stop Words網站[3]所列出之文字。之後，利用Rubryx2.0來統計符合POS規則的特徵之頻率。並且建構詞彙-文件矩陣，如表三：

表三、詞彙-文件矩陣之範例

Keywords / document	feel good	film noir	great deal	…
P001	0	0	2	…
P002	1	0	0	…
P003	0	1	0	…
P004	1	1	0	…
P005	0			

橫向表示本研究所擷取之關鍵字，縱向則表示文章內容編號。每一個細胞則代表了，此關鍵字在這篇文章中出現的次數。例如：圓圈部份表示 great deal 這個關鍵字，出現在編號 P001 這篇文章中，一共出現了兩次。

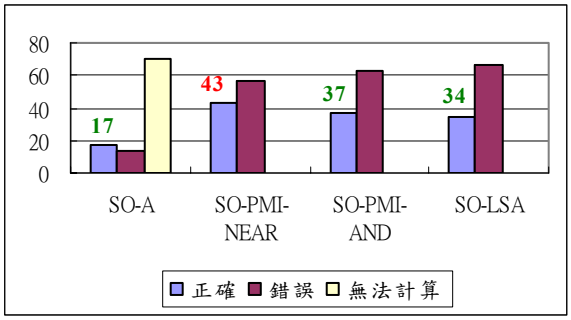
4.2實驗結果

4.2.1 傳統語意導向指標

本研究將以 100 篇電影評論為實驗文集。實驗文集中有 31 為負向情緒之內容，69 筆為正向情緒之內容。在本實驗模型中，我們從 Rubryx2.0 軟體統計出來的詞頻集合中，挑選出頻率較高的正向情緒關鍵字和負向情緒關鍵字，並使構成正向情緒集合和負向情緒集合，分別為 Pwords={feel good, good movie, good performance}、Nwords={film noir, long time, old fashioned}。

表四、傳統語意導向指標之精確度

Performance Index	精確率	錯誤率	無法辨識
SO-A	17%	13%	70%
SO-PMI-NEAR	**43%**	57%	-
SO-PMI-AND	37%	63%	-
SO-LSA	34%	66%	-



圖四、傳統語意導向指標之分類結果

由表四和圖四中，我們可以發現在傳統四種語

1.http://www.cs.cornell.edu/people/pabo/movie-review-data/
2. http://www.sowsoft.com/rubryx/
3.http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

意導向指標中，以 SO-PMI-NEAR 之計算結果最佳，可以獲得 43%的精確率。

4.2.2 基於類神經網路之語意導向指標

過去我們也曾提出一個基於類神經網路之語意導向指標。主要目的為提升傳統語意導向指標的執行效率。由過去的實驗我們可以由表五得知，在基於類神經網路之語意導向指標的方法中，若訓練集合為 60%資料量(40%資料當成測試樣本)、隱藏層為 1 層、神經元個數為 4、學習率為 0.1 及學習次數為 50 時，可以獲得 70%的正確率。依整體看來平均正確率大約有 69%，大幅地提升了傳統語意導向方法的執行效率。

表五、基於類神經網路之語意導向指標之實驗結果

訓練樣本資料量 / 績效與參數設定	50%	60%	70%	80%	90%
正確率	69%	**70%**	69%	69%	69%
錯誤率	31%	**30%**	31%	31%	31%
網路架構	4-4-1	**4-4-1**	4-3-1	4-3-1	4-3-1
學習率	0.01 /0.1	**0.1**	0.01	0.01	0.1 /0.01
學習次數	100 /100	**50**	100	100	45 /100

註：訓練樣本資料量 60%，表示利用 60%資料當作訓練樣本，40%資料則為測試樣本

除此之外，我們也將傳統語意導向藉由 -1、0 與 1，做為基於 NN 的正規化語意導向指標以及將原始詞彙-文件矩陣，透過倒傳遞類神經網路做分類。表六為基於 NN 的正規化語意導向指標之實驗結果，若在訓練集合為 80%、隱藏層為 1 層、神經元個數為 4、學習率為 0.4 及學習次數為 8 時，同樣可以獲得 70%的執行正確率。而表七則為詞彙-文件矩陣的分類結果，當訓練集合為 60%與 90%、隱藏層皆為 1 層、神經元個數分別為 4 個與 3 個、學習率皆為 0.01 以及學習次數分別為 69 與 70 時，皆可以獲得 69%的精確率。

表六、基於類神經網路的正規化語意導向指標之實驗結果

訓練樣本資料量 / 績效與參數設定	50%	60%	70%	80%	90%
正確率	69%	69%	69%	**70%**	69%
錯誤率	31%	31%	31%	**30%**	31%
網路架構	4-4-1	4-3-1	4-4-1	**4-4-1**	4-2-1
學習率	0.1 /0.5	1	0.01	**0.4**	0.01
學習次數	30 /5	10	100	**8**	100

註：訓練樣本資料量 60%，表示利用 60%資料當作訓練樣本，40%資料則為測試樣本

表七、基於類神經網路的詞彙-文件分類之結果

訓練樣本資料量 / 績效與參數設定	50%	60%	70%	80%	90%
正確率	68%	**69%**	68%	64%	**69%**
錯誤率	32%	**31%**	32%	36%	**31%**
網路架構	4-3-1	**4-4-1**	4-3-1	4-2-1	**4-3-1**
學習率	0.01	**0.01**	0.01	0.1	**0.01**
學習次數	70	**69**	100	45	**70**

註：訓練樣本資料量 60%，表示利用 60%資料當作訓練樣本，40%資料則為測試樣本

4.2.3 基於支援向量機之語意導向指標

而 Joachims(1998)和 Yang et al.(1999)曾經利用支援向量機於主題內容分類的實驗中，實驗執行之精確度皆可達到 85%以上。因此，本研究將提出一個基於 SVM 為主之語意導向指標。希望除了提升傳統語意導向指標之執行效率外，亦可提升以機器學習為主的語意導向指標的執行成效。表八為基於 SVM 之語意導向指標執行結果。當訓練集合為 90%，最佳的參數 C 為 0.03125 及參數 γ 為 0.007813，測試樣本為 10 筆資料時，可以獲得 100%的高正確率。依整體看來平均正確率大約有 77%。由此可證明，本研究所提出之基於 SVM 語意導向指標確實可以提升機器學習方法的語意導向指標之執行效率。

表八、基於 SVM 之語意導向指標之實驗結果

績效與參數設定 / 訓練樣本資料量	參數 C	參數 γ	正確率	錯誤率
50%	0.03125	0.007813	68%	32%
60%	0.03125	0.007813	67%	33%
70%	0.03125	0.007813	70%	30%
80%	0.03125	0.007813	80%	20%
90%	**0.03125**	**0.007813**	**100%**	**0%**

除此之外，我們也將傳統語意導向藉由 -1、0 與 1，做為基於 SVM 的正規化語意導向指標以及將原始詞彙-文件矩陣，透過支援向量機做分類。表九為基於 SVM 的正規化語意導向指標的實驗結果，當訓練集合為 90%，最佳的參數 C 為 0.5 及參數 γ 為 2，測試樣本為 10 筆資料時，可以獲得 80%的正確率。而表十為基於 SVM 的詞彙-文件矩陣分類結果，當訓練集合為 70%與 80%，最佳的參數 C 為 0.0312 及參數 γ 為 0.007813，測試樣本分別為 30 筆與 20 筆資料時，皆可以獲得 75%的正確率。

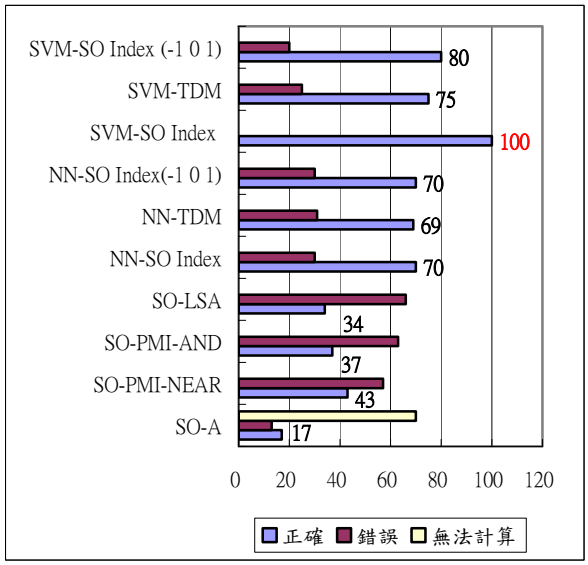
表九、基於 SVM 的正規化語意導向指標之實驗結果

績效與參數設定 / 訓練樣本資料量	參數 C	參數 γ	正確率	錯誤率
50%	32768	0.001953	64%	36%
60%	0.03125	0.007813	72.50%	27.50%
70%	0.03125	0.007813	70%	30%
80%	0.03125	0.007813	70%	30%
90%	**0.5**	**2**	**80%**	**20%**

表十、基於 SVM 的詞彙-文件矩陣分類之結果

績效與參數設定 / 訓練樣本資料量	參數 C	參數 γ	正確率	錯誤率
50%	0.03125	0.007813	66%	34%
60%	0.03125	0.007813	70%	30%
70%	**0.03125**	**0.007813**	**75%**	**25%**
80%	**0.03125**	**0.007813**	**75%**	**25%**
90%	0.03125	0.007813	70%	30%

圖五為基於 SVM 之 SO 指標以及其它相關的實驗之結果比較圖。由圖可知，本研究所提出之基於 SVM 之語意導向指標，不論與非監督的傳統語意導向指標或為與機器學習相關的基於 NN 之 SO 指標相比，正確率皆爲較精準的一個指標。



圖五、基於 SVM 之 SO 指標與其它實驗之結果比較

5.結論

本研究所提出之基於 SVM 之語意導向指標，當訓練樣本為 90%時可獲得 100%的正確率。因此，經由實驗證實本研究所提出之方法，不但可以改善傳統語意導向指標之分類結果外，也可提升了基於機器學習之語意導向指標中的類神經指標之效能。但因為本研究，目前只利用 100 筆的電影評論，實踐本研究所提出之想法。因此，未來我們除了增加實驗文集之外，還必須擴大實驗文集之範疇，以驗証本文所提出之方法是否適用於各種領域的部落格內容中。

另外，本研究是結合機器學習及資訊檢索技術的優點，以提出基於 SVM 之語意導向指標。非監督的語意導向方法，雖然精確度較機器學習低，但具備了即時使用的優勢；機器學習雖然精確度較高，但卻必須花費大量的時間訓練分類模型。因此，未來我們可以試著改良非監督語意導向相關方法，除了希望可以改善傳統語意導向方法的執行效能之外，並且以改

善花費大量時間訓練分類模型為目標。

6.致謝

本研究受到國科會計畫(契約編號 NSC 96-2416-H-324 -003 -MY2)部分贊助，作者在此表達感謝之意。

參考文獻

[1] Abbasi, A., Chen, H., Thoms, S. and Fu, T., "Affect analysis of web forums and blogs using correlation ensemblesT," ***IEEE Transactions on Knowledge and Data Engineering,*** vol. 20, pp.1168-1180, 2008.

[2] Bartell, B.T., Cottrell, G.W., and Belew, R.K. "Latent semantic indexing is an optimal special case of multidimensional scaling," ***Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,*** pp. 161-167, 1992.

[3] Chen, Y., Tsai, F. S. and Chan, K. L. "Machine learning techniques for business blog search and mining," ***Expert Systems with Applications: An International Journal,*** Vol. 35, pp. 581-590, 2008.

[4] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A., "Indexing by latent semantic analysis," ***Journal of the Society for Information Science,*** Vol. 41, pp. 391-407, 1990.

[5] Hatzivassiloglou, V. and McKeown, K. R., "Predicting the semantic orientation of adjectives," ***presented at Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL,*** pp.174-181, 1997.

[6] Herring, S.C., Scheidt, L.A., Bonus, S. and Wright, E**.,** **"**Bridging the gap: a genre analysis of Weblogs," ***Proceedings of the 37th Hawaii International Conference on System Sciences***, 2004.

[7] Joachims, T., "Text categorization with support vector machines: Learning with many relevant features," ***Proceedings of 10th European Conference on Machine-learning,*** pp. 21-24 (pp. 137-142), 1998.

[8] Landauer, T.K. and Dumais, S.T., "A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge," ***Psychological Review***, pp. 211-240, 1997.

[9] Leshed, G. and Kaye, J., **"**Understanding how bloggers feel: recognizing affect in blog posts**,"** ***Conference on Human Factors in Computing Systems,*** pp. 1019-1024, 2006.

[10] Martin, J., "Blogging for dollars," ***FSB: Fortune Small Business***, Vol. 15(10), pp. 88-92, 2005.

[11] Mishne, G. and Rijke, M. D., "Capturing Global Mood Levels using Blog Posts," **Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to AnalysingWeblogs,** pp. 145-152. 2006.

[12] Murphy, C., "Blogging: Waste of time or corporate tool?" Retrieved August 12, 2006, from http://www.personneltoday.-com/Articles/2006/03/21/34506/blogging-waste-of-time-orcorporate- tool.html, 2006.

[13] Pang, B., Lee, L. and Vaithyanathan, S., "Thumbs up? Sentiment classification using machine learning techniques," ***presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002)***, pp.79-86, 2002.

[14] Singh, T., Veron-Jackson, L., Cullinane, J., "Blogging: A new play in your marketing game plan," ***Business Horizons,*** Vol. 51, pp. 281-292, 2008.

[15] Subasic, P. and Huettner, A**.,** **"**Affect analysis of text using fuzzy semantic typing**,"** ***The Ninth IEEE International Conference of Fuzzy Systems, Vol. 2, pp.*** 647-652, 2000.

[16] Turney, P. D. and Littman, M. L., "Measuring praise and criticism: inference of semantic orientation from association," ***ACM Transactions on Information Systems*****,** vol. 21, pp. 315-346, 2003.

[17] Turney, P. D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," ***Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, Pennsylvania, USA,*** pp.417-424, 2002.

[18] Winer, D., "The history of weblogs," http://newhome.weblogs.comhistory of Weblogs, 2002.

[19] Wu, C. H. Chuang, Z. J. and Y. C. Lin, **"**Emotion recognition from text using semantic labels and separable mixture

models," ***ACM Transactions on Asian Language Information Processing,*** Vol. 5, pp. 165-183, 2006.

[20] Yang, Y. and Liu, X. "A re-examination of text categorization methods," ***Proceedings of 22nd ACM International Conference on Research and Development in Information Retrieval***, pp. 16–19 (pp. 42–49), 1999.

[21] Yap, R. Muirhead, B. and Keefer, J., "Blog RUBRIC: Designing your Business Blog," ***International Journal of Instructional Technology and Distance Learning***, Vol. 2, pp. 53-60, 2005.

[22] Smola, A. J., "Learning with Kernels," ***PhD Thesis, GMD, Birlinghoven, Germany,*** 1998.

[23] Chaovalit, P. and Zhou, L., "Movie Review Mining: a Comparison between Supervised and Unsupervised," ***Proceedings of the 38th Hawaii International Conference on System Sciences,*** pp. 112c- 112c, 2005.

[24] Landauer, T. K. AND Dumais, S. T., "A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge," ***Psychological Review,*** 104, 211-240, 1997.

[25] Golub, G. H. and Van LOAN, C. F. "*Matrix Computation,"*. ***Third edition. Johns Hopkins University Press, Baltimore, MD***, 1996.

[26] Burges, C. J. C., "A Tutorial on Support Vector Machines for Pattern Recognition" ***Data Mining and Knowledge Discovery***, pp. 1-47, 1998.

[27] Hu, M. and Liu, B., "Mining and summarizing customer reviews," ***The 2004 SIGKDD***, 168-177, 2004.