提升語意分類績效之研究

陳隆昇
朝陽科技大學資訊管理系
lschen@cyut.edu.tw

張家偉
朝陽科技大學資訊管理系
s9814628@cyut.edu.tw

摘要

隨著文字為主的溝通平台，例如：部落格、微博、推特等的快速發展，網路使用者對購買產品或服務的評價可急速在網路空間擴散，並直接影響其他顧客的購買意願及品牌印象，特別是負面評價。因此如何偵測出使用者的語意已成為新興的研究議題。語意分類主要是區分文字評論成正向或負向的意見，並提供企業產品改善依據或對顧客抱怨之因應態度。近來，機器學習方法結合特徵選取、及特徵權重的方法已被證實可以有效處理語意分類的問題。然而，在相關研究中，並未提及特徵權重、特徵選取、及機器學習方法最佳組合為何。因此，本研究主要目的在找出機器學習方法、特徵選取方法與特徵權重的最佳組合，以提升文字語意的分類績效。

關鍵詞：特徵選取、特徵權重、語意分類、文字探勘

1. 前言

隨著網際網路的普遍，愈來愈多人使用基於文本(text)的溝通工具，例如：部落格(Blogs)、推特(Twitter)等，以表達他們對於一個產品或服務的意見和評論。這些網路評論會對消費者的購買行為產生影響。根據電子商務服務公司Channel Advisor在2010年針對消費者購物習慣的調查報告中指出，有92%的消費者表示當他們考慮購買一個產品時，會先上網閱讀關於此產品的評論，其中又有46%的人表示產品評論的內容會對他們的購買行為產生影響，而有43%的人則會由於負面的產品評論因而放棄購買此產品。由此可知，負面的評論有可能會降低消費者的購買意願，更會對企業造成巨大的損失。

此外，隨著網路評論數量的增加，人們愈來愈難去得知關於一個產品主要的意見(Zhang et al., 2008)，加上網路評論常常是非結構化的、主觀的，難以在短時間內理解(Chaovalit and Zhou, 2005)。因此，如何從大量的網路評論中有效地辨識顧客所表達的語意，變成一項新興的研究議題。

語意分類(Sentiment Classification)技術是一個新近發展的網站探勘技術，可以進行語意或意見上的分析(Liu et al., 2005)。語意分類主要是區分文字評論成正向或負向的意見(Ye et al., 2009)，並提供企業產品改善依據或對顧客抱怨之因應態度。近來，機器學習方法已經被證實可以有效地處理語意分類的問題(Pang et al., 2002; Tan and Zhang, 2008; Zhang et al., 2008; O'Keefe and Koprinska, 2008; Ye et al., 2009; Bai, 2010)。但是網路評論的數量每天都在增加，文本資料的特徵維度也不斷在提高，也使得機器學習方法的效能下降。因此，如何有效地選取有用的特徵，降低特徵空間的維度以改善分類的效能變得相當重要。已經有許多研究使用特徵選取方法進行網路評論的語意分類，例如：Ye et al. (2009)使用Information gain(IG)做為特徵選取進行旅遊評論的語意分類、O'Keefe 和 Koprinska(2009)使用類別比例差異方法(Categorical Proportional Difference, CPD)進行電影評論的語意分類，實驗結果皆顯示特徵選取方法的導入可以改善語意分類的效能。此外，當使用機器學習方法訓練分類模型時，每一份文件會以特徵向量的形式被表示(Zhang et al., 2008)，也就是計算特徵權重。在語意分類的相關研究中已經有許多的特徵權重方法被使用，例如：Term frequency (TF) (Pang et al., 2002; Na et al., 2005)、Term frequency-inverse

document frequency (TF-IDF) (Tan and Zhang, 2008; Zhang et al., 2008)、Term presence (TP) (Bai, 2010; Abbasi et al., 2007)。

然而，在相關研究中，並未提及特徵權重、特徵選取、及機器學習方法最佳組合為何。因此，本研究主要目的在找出機器學習方法、徵選取方法與特徵權重的最佳組合，以提升文字語意的分類績效。本研究會以一個實際的產品評論網站的資料來評估與驗證所有方法的有效性，並做出相關建議。

2. 文獻探討

2.1 語意分類

語意分類已經被廣泛地應用在許多領域，例如：產品的比較、產品評論的彙整和意見的探勘(Tan et al., 2009)。企業可以藉由語意分類的結果去得知一個產品被接受的程度，以進一步改善產品的品質(Prabowo and Thelwall, 2009)。商業部落格可以藉由語意分類所提供的資訊，去妥善回應讀者的訊息(Lee et al., 2002)。顧客可以藉由語意分類的結果，去獲得關於一個產品主要的意見，以決定是否要購買此產品(Zhang et al., 2008)。此外，已經有許多研究進行了文本資料的語意分類，而這些研究所使用的方法主要可以分成兩類(Tan and Zhang, 2008)。第一類是機器學習方法(Machine Learning)，主要是基於文本的語意資料中各種不同文字出現的頻率(即詞彙文件矩陣)，去訓練一個語意分類器，再以此分類器來辨識文字語意。第二類是語意導向方法(Semantic Orientation)，其主要作法是先建立正向與負向的詞彙集，然後分別計算某一文件與正向或負向詞彙集的關係分數，以判斷該文件之語意導向。

文獻上指出機器學習方法的分類效能通常較佳，但是需要花時間訓練分類模型，並且需要額外的類別資訊才能進行。而語意導向方法雖然能夠快速地解決語意分類的問題，但是其分類準確率通常較低(Chaovalit and Zhou, 2005)。因為機器學習方法有較佳的績效，因此本研究僅以機器學習方法做為研究標的。

2.2 特徵選取方法

特徵選取是指從文件中選取能夠表現該文件的特徵(Wang et al., 2010)。藉由特徵選取可以降低特徵空間的維度、減少計算時間與成本，並且移除可能的雜訊和改善分類的效能(Chen et al., 2009; Li et al., 2007; Polat and Gunes, 2009; Karabatak and Ince, 2009)。

在語意分類的相關研究中許多特徵選取方法已經被使用。如表 1 所示，大致上可以分為三類，分別是(1)基於文件中詞語出現頻率的方法、(2)基於統計理論的方法、與(3)基於 Part of speech(POS)標籤的方法。基於文件中詞語頻率的方法是計算方式較為簡單的方法，而特徵頻率(Feature frequency, FF)是普遍被使用的特徵選取方法，其優點是容易計算(Tang and Zhang, 2008)。但是與基於統計理論的方法相比，FF 的分類效能通常較差(Tan and Zhang, 2008)。此外，也有學者針對基於文件中詞語頻率的方法進行改善，例如：Simeon 和 Hilderman(2008)提出類別差異(Categorical Proportion Difference, CPD)方法，使用詞語的文件頻率來進行計算。雖然基於統計理論的方法較能有效地選取特徵，有較佳的分類效能，但是需要花費較多的計算成本(O'Keefe & Koprinska, 2009)，其中 IG 是普遍被使用的方法。另外，Tan 和 Zhang (2008)使用 IG 進行網路產品評論的語意分類，實驗結果顯示其分類效能較優於卡方分配(Chi square, CHI)與互斥資訊(Mutual Information, MI)。而基於 POS 標籤的方法根據 Li et al.(2007)、Na et al.(2005)和 Pang et al.(2002)之研究，皆證實無法有效地改善語意分類的效能。在本研究中，我們使用特徵頻率(FF)、類別比例差異(CPD)、IG 做為特徵選取方法，進行網路評論的語意分類，以下將分別介紹三種方法：

(1) 特徵頻率(FF)

特徵頻率是指一個特徵在所有文件出現的頻率，是一種普遍使用在文本分類的特徵選取方法 (Li et al., 2007; Na et al., 2005; Pang et al., 2002; Tan and

Zhang, 2008)。一個特徵的特徵頻率可以透過下列公式(1)來表示：

$$FF = \sum_{j}^{n} (tf)_j \quad (1)$$

其中$(tf)_j$表示此特徵在第 j 份文件出現的頻率，n 表示所有文件的總數。

(2) 類別比例差異(CPD)

CPD 方法是由 Simeon 和 Hilderman(2008)所提出的一個特徵選取方法，應用於多類別的文本分類。O'Keefe 和 Koprinska(2009)將此方法應用於兩類別的語意分類之研究。此方法主要是分別計算一個特徵的正向文件頻率(Positive DF)與負向文件頻率(Negative DF)，然後再計算一個特徵在兩個類別之間分佈比例的差異。一個特徵的 CPD 值可以透過公式(2)來表示：

$$CPD = \frac{Positive\ DF - Negative\ DF}{Positive\ DF + Negative\ DF} \quad (2)$$

由公式(2)可以得知 CPD 值會介於 0 到 1 之間，如過一個特徵僅出現在單一類別(正向文件或是負向文件)，可以得到 CPD 值等於 1，這對於分類而言是有用的特徵；相反地，如果一個特徵出現在兩個類別的文件頻率相同，則可以得到 CPD 值等於 0，這表示該特徵對於分類無法提供有用的資訊。

(3) Information gain(IG)

主要是藉由熵(Entropy)的概念，量測一個特徵有出現在一份文件中，或是沒有出現在一份文件中時所獲得的資訊量去預測類別(Tan and Zhang, 2008; Simeon and Hilderman, 2008)，其計算方式如下：

$$\begin{aligned} IG(t) = & -\sum_{i=1}^{|C|} P(c_i)\log(c_i) \\ & + P(t)\sum_{i=1}^{|C|} P(c_i \mid t)\log P(c_i \mid t) \\ & + P(\bar{t})\sum_{i=1}^{|C|} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \end{aligned} \quad (3)$$

其中 $P(c_i)$表示類別 c_i 出現的機率，$P(t)$表示文字 t 出現的機率，$P(\bar{t})$表示文字 t 沒有出現的機率。

2.3 特徵權重方法

在資訊檢索(Information retrieval)的領域，特徵權重主要被用來表示一個特徵在檢索過程中的有用性(Aizawa, 2003)。當進行文本資料的分類時，每一份文件會以特徵向量的形式被表示(Zhang et al., 2008)，也就是建構詞彙文件矩陣(Term-Document Matrix, TDM)。在詞彙文件矩陣中，每一個特徵向量代表一個詞語在一份文件中的權重。在文本分類的相關研究中已經有許多特徵權重方法被使用，例如：

(1) Term Frequency (TF)

TF 是使用一個詞語在一份文件中出現的次數來表示權重，它主要是衡量一個詞語對於一份文件的重要性，又可以稱為「局部詞彙權重(Local Term Weight)」(Tian and Tong, 2010)。

(2) Inverse Document Frequency (IDF)

IDF 是藉由一個詞語的文件頻率之倒數來計算權重，它主要是衡量一個詞語在所有文件出現的普遍程度，又可以稱為「整體詞彙權重(Global Term Weight)」(Tian & Tong, 2010)。一個詞語 t 的 IDF 權重可以透過下列公式(4)來表示：

$$IDF = log\frac{N}{m_t} \quad (4)$$

在公式(4)中，N 表示所有文件的總數，m_t 表示包含特徵 t 的文件總數。

(3) Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF 是結合 TF 權重與 IDF 權重，是一個普遍使用在資訊檢索領域的權重方法(Aizawa, 2003; Singhal, 2001; Tian and Tong, 2010)。TF-IDF 可以透過下列公式(5)來計算：

$$TF - IDF = TF \times IDF \quad (5)$$

(4) Term presence (TP)

TP 是以一個詞語是否出現於一份文件中來表示權重，如果有出現就以 1 表示，沒有出現就以 0 表示。Pang et al. (2002) 首先將 TP 權重使用於兩類別的語意分類之研究，並且與 TF 權重進行比較。其實驗結果顯示，使用 TP 權重進行語意分類之效能較優於 TF 權重。

為了探討特徵權重方法對於語意分類效能的影響，本研究分別使用 TF、IDF、TF-IDF、TP 等方法來計算特徵權重，進行網路評論的語意分類。

表 1 特徵選取方法分類

分類	方法
基於文件中詞語頻率的方法	特徵頻率(Na et al., 2005; Pang et al., 2002)
	類別比例差異(O'Keefe and Koprinska, 2009)
基於統計理論的方法	卡方分配(Tian and Tong, 2010; Zhang et al., 2007)
	Information gain (Abbasi et al., 2008; Bai, 2010; Ye et al., 2009; Zhang et al., 2011)
	Mutual information (Tan and Zhang, 2008)
基於 Part of speech(POS)標籤的方法	POS (Li et al., 2007; Mullen and Collier, 2004)

2.4 機器學習方法

已經有許多研究使用機器學習方法來處理語意分類的問題，詳如表 2。其中支撐向量機(Support Vector Machines, SVM)普遍有較佳的分類效能。此外，Khan et al. (2009)之研究隨機選取了 336 篇使用機器學習方法之相關文獻進行統計，其結果如圖 1，其顯示 SVM 在近年來逐漸成為普遍使用的機器學習方法。因此，本研究主要使用 SVM 做為機器學習方法，進行網路評論的語意分類。

SVM 是由 Vapnik (1995) 根據統計學習理論中結構風險最小化(structural risk minimization)原則所提出的機器學習方法。SVM 可以用來處理兩類別或是多類別資料分類的問題。在語意分類的相關研究中，SVM 主要用來處理兩類別分類的問題，藉由尋求一個超平面(hyperplane)有最大的邊界(Margin)來區分兩類別的資料(可參考圖 2)。

表 2 機器學習方法分類效能比較

作者	實驗資料	機器學習方法	分類結果(準確率)
Pang et al. (2002)	電影評論	SVM	82.90%
		Naive Bayes	78.70%
		Maximum Entropy	77.70%
Tang and Zhang (2008)	教育、電影、房屋評論	SVM	90.43%
		Naive Bayes	88.82%
		KNN	87.30%
		Centroid	86.81%
		Winnow	89.96%
Zhang et al. (2008)	書籍、音樂、電影、電子商務、數位產品評論	SVM	80.26%
		Naive Bayes	79.08%
		Decision tree	70.52%
O'Keefe and Koprinska (2008)	電影評論	SVM	87.15%
		Naive Bayes	81.50%
Ye et al. (2009)	旅遊評論	SVM	86.06%
		Naive Bayes	80.71%
		N-gram model	84.05%
Bai (2010)	電影評論	SVM	84.07%
		Naive Bayes	73.10%
		Voted perc.	72.30%
		Maximum Entropy	79.43%

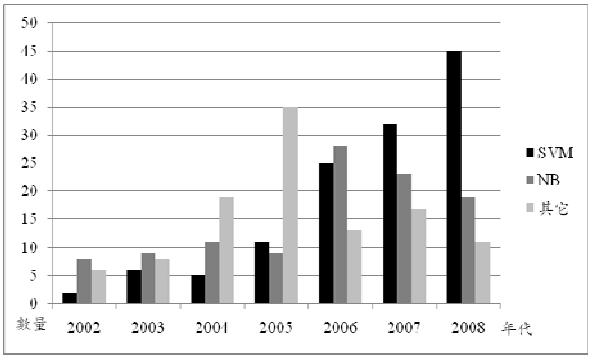

圖 1 文獻使用機器學習方法之統計(Khan et al., 2009)

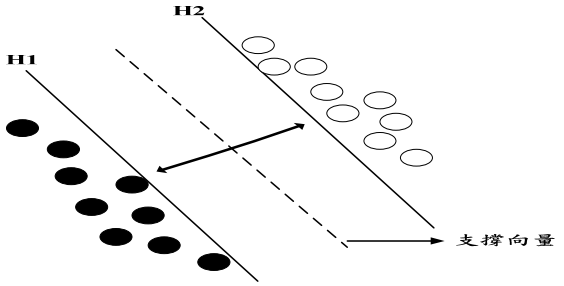

圖 2 SVM 區分超平面示意圖(資料來源：修改自 Unler et al., 2010)

3. 研究方法

此章節將說明本研究方法之流程,可以分為四個步驟：

步驟一： 資料前處理

在蒐集實驗文集之後我們以 unigram 做為一個特徵的基本單位，接著使用一個 stop words 列表(http://snowball.tartarus.org/algorithms/english/stop.txt)去刪除文章中的冗詞、贅字，然後建立一個以 unigram 為單位的特徵集合。另外，我們也將此只做資料前處理的特徵集合，使用四種特徵權重方法計算權重後進行分類，並將此分類結果做為基準，探討特徵選取方法的導入與結合各種不同特徵權重方法之後對於分類效能的影響。此外，本研究使用五折交叉驗證(five cross-validation)去建立訓練資料集合與測試資料集合。

步驟二： 特徵選取

我們分別使用特徵頻率、類別比例差異、Information gain 等方法去計算每一個特徵的重要度，然後根據重要度將特徵進行排序。此外，為了在相同的特徵維度下進行比較，我們使用預先定義的幾個特徵維度來選取特徵(例如：1000、700、400…等)。以下將以一個簡單的例子來說明：

表 3 中有三個特徵，我們將分別使用三種特徵選取方法計算其重要度，而預先定義的維度為 2。

(1) FF 方法

以特徵頻率來計算重要度，重要度排序：特徵 B > 特徵 C > 特徵 A。

選取的特徵：特徵 B、特徵 C。

(2) CPD 方法

特徵 A：(3-1)/(3+1)=0.5

特徵 B：(3-2)/(3+2)=0.2

特徵 C：(4-0)/(4+0)=1

重要度排序：

特徵 C > 特徵 A > 特徵 B。

選取的特徵：特徵 C、特徵 A。

(3) IG 方法

特徵 A：

-(-1)+(-0.311)+(-0.5)=0.189

特徵 B：

-(-1)+(-0.442)+(-0.529)=0.029

特徵 C：

-(-1)+(0)+(-0.39)=0.61

重要度排序：

特徵 C > 特徵 A > 特徵 B。

選取的特徵：特徵 C、特徵 A。

表 3 特徵選取說明例

特徵	特徵頻率	正向文件頻率	負向文件頻率
特徵 A	6	3	1
特徵 B	15	3	2
特徵 C	8	0	4
文件總數：10 正向文件總數：5 負向文件總數：5			

步驟三： 計算特徵權重

將所選取出來的特徵分別使用四種特徵權重方法(TF、IDF、TF-IDF、TP)計算特徵權重，建立詞彙文件矩陣。以下將以一個簡單的例子來說明：

表 4 詞彙文件矩陣說明例

	特徵 A	特徵 B	特徵 C	類別
文件 1	0	4	0	1
文件 2	2	2	0	1
文件 3	1	0	0	1
文件 4	1	0	0	1
文件 5	0	2	0	1
文件 6	0	0	2	-1
文件 7	2	4	3	-1
文件 8	0	3	2	-1
文件 9	0	0	1	-1
文件 10	0	0	0	-1

上表 4 為表 3 中的三個特徵使用 TF 方法計算特徵權重所建立之詞彙文件矩陣。橫軸代表特徵，縱軸代表每一份文件，最後一欄為文件所對應之類別。在矩陣中的每一個數值，則代表每一個特徵在每一份文件中的權重。同理，我們可以分別用 TP、TF-IDF 與 IDF 權重來分別構建不同之詞彙文件矩陣。

步驟四：機器學習訓練

將訓練資料集合輸入 SVM 分類器訓練分類模型，然後將測試資料集合輸入所建立的分類模型去測試分類效能。

4. 實驗結果

4.1 實驗資料與工具

本研究以網路評論作為研究案例，從 Review Centre 網站下載 MP3 相關的產品評論做為實驗資料，基本資料如表 5 所示。在 Review Centre 網站中以 5 個星等將評論做評比，我們將 4 或 5 個星等的評論視為正向的評論，1 或 2 個星等的評論視為負向的評論，3 個星等的評論則視為中立的評論而不採用。本研究總共下載了 MP3 產品評論 400 筆，包含了正向評論 200 筆與負向評論 200 筆。另外，我們使用 Chang 和 Lin (2001)所開發的 LIBSVM 做為 SVM 分類器，核心函數為預設的 Radial Basis Function (RBF)函數。

表 5 實驗資料

實驗資料	資料數量	特徵數量	資料來源
MP3 產品評論	400	1382	http://www.reviewcentre.com/

4.2 實驗結果

首先，原始資料僅進行前處理，未使用任何特徵選取方法，然後使用四種特徵權重方法進行實驗，以做為比較基礎，其結果如表 6 所示。從表 6 可以得知，當未使用特徵選取方法進行分類時，使用 TP 權重在 SVM 分類器中有較佳的分類績效表現。

接下來我們導入三種特徵選取方法 (FF、CPD、IG)，探討 SVM 分類器分別在 TP、TF、IDF、TF-IDF 四種特徵權重下的表現。其結果分別羅列於圖 3 至圖 6。

在以 TP 及 TF 為特徵權重下，我們可以發現到，CPD 特徵選取法在使用較高的特徵維度有較佳的分類結果，但是隨著特徵維度的遞減，其分類效能則明顯地下降。但 IG 方法和 FF 方法之分類效能則是較不受到維度遞減的影響呈現就穩定的變動。其中 IG 方法又明顯優於 FF 方法。尤其是隨著特徵維度的減少，IG 方法普遍有較佳的分類結果。

當以 TF-IDF 及 IDF 為權重表示法時，IG 法則展現了過人的績效；在維度從 1382 縮減至 1000 及 700 時，CPD 還能與其維持不遠的績效。然而，當維度再往下縮減時，IG 則明顯優於 CPD 與 FF。

此外，在表 7 中，我們亦列出了三種特徵選取方法之最佳分類結果。從此表中，我們發現到 FF 方法使用 TP 權重時有最佳的分類效能，CPD 方法則是使用 TF-IDF 權重時有最佳的分類效能，而 IG 方法使用 IDF 權重時有最佳的分類效能，並且優於其他兩種方法。

表 6 資料未使用特徵選取方法之分類結果

實驗 \ 權重	TP	TF	IDF	TF-IDF
Fold-1(%)	83.75	80.00	78.75	81.25
Fold-2(%)	97.50	92.50	93.75	93.75
Fold-3(%)	80.00	80.00	80.00	78.75
Fold-4(%)	82.50	81.25	87.50	83.75
Fold-5(%)	76.25	76.25	73.75	70.00
Mean	84.00	82.00	82.75	81.50
SD	8.07	6.16	7.88	8.59

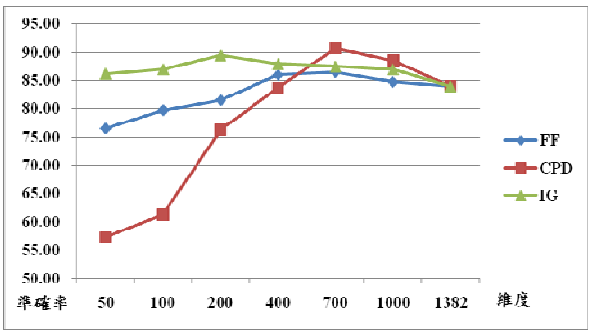

圖 3 TP 權重之分類結果

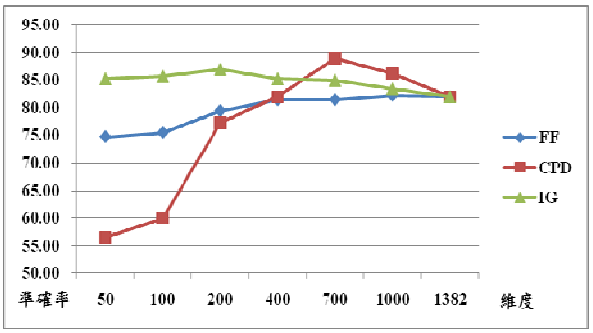

圖 4 TF 權重之分類結果

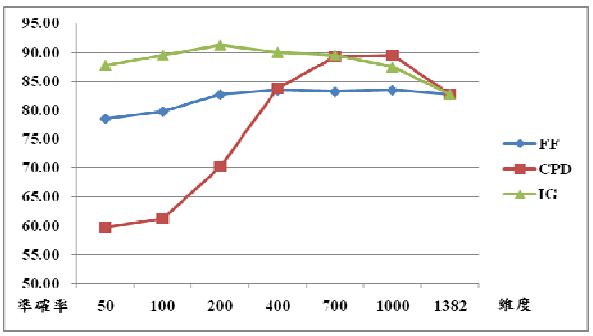

圖 5 IDF 權重之分類結果

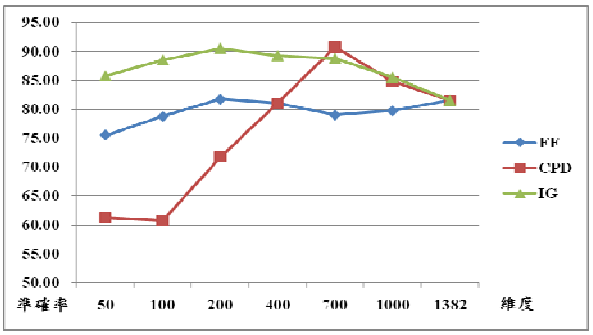

圖 6 TF-IDF 權重之分類結果

表 7 三種特徵選取方法之最佳分類結果

權重 / 方法	TP	TF	IDF	TF-IDF
FF	86.50	82.25	83.50	81.75
CPD	90.50	89.00	89.50	90.75
IG	89.50	87.00	91.25	90.50

5. 結論與未來研究方向

本研究使用三種特徵選取方法與四種特徵權重方法，並結合 SVM 分類器進行 MP3 產品評論的語意分類。從實驗結果可以做出幾點結論。首先，CPD 方法在使用較高的維度進行分類時有較佳的分類效能，但是其分類效能容易受到維度遞減的影響而明顯下降。其次，FF 方法和 IG 方法之分類效能較不受到維度遞減的影響。最後，當 FF 方法使用 TP 權重時有最佳的分類效能，而 CPD 方法使用 TF-IDF 權重時有最佳的分類效能。IG 方法則是使用 IDF 權重時有最佳的分類效能，並且當使用較少的維度進行分類時，也有最佳的分類準確率。

由於網路評論的數量每天都不斷在增加，因此未來可以嘗試增加實驗文集的數量，或是對不同類型的語意資料進行分析，或許可以得到更穩健的結論。此外，影音等其他多媒體資料亦越來越普遍，後續的研究可以針對其他類型的評論進行實驗。

致謝

本研究受到國科會計畫(契約編號 NSC 98-2410-H-324-007-MY2)部分贊助，作者在此表達感謝之意。

參考文獻

[1] A. Aizawa (2003), *An Information-theoretic Perspective of TF-IDF Measures*, Information Processing and Management, Vol. 39, No. 1, pp. 45-65.

[2] A. Singhal (2001), *Modern Information Retrieval: A Brief Overview*, IEEE Data Engineering Bulletin, Vol. 24, No. 4, pp. 35-43.

[3] A. Unler, A. Murat, and R.B. Chinnam (2010), *Mr2PSO: A Maximum Relevance Minimum Redundancy Feature Selection Method Based on Swarm Intelligence for Support Vector Machine Classification*, Information Sciences, doi:10.1016/j.ins.2010.05.037.

[4] B. Li, S. Xu, and J. Zhang (2007), *Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments*, Proceedings of the 45th Annual Southeast Regional Conference, pp. 94-99.

[5]B. Liu, M. Hu, and J. Cheng (2005), *Opinion Observer: Analyzing and Comparing Opinions on the Web*, Proceedings of the 14th International Conference on World Wide Web.

[6]B. Pang, L. Lee, and S. Vaithyanathan (2002), *Thumbs up?: Sentiment Classification Using Machine Learning Techniques*, Annual Meeting of the ACL Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp.79-86.

[7]C.C. Chang and C.J. Lin (2001), *LIBSVM: a Library for Support Vector Machines, Software*, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8]ChannelAdvisor Corporation (2010), *Through the Eyes of the Consumer: 2010 Consumer Shopping Habits Survey*, http://www.channeladvisor.com/

[9]C.M. Lee, S.S. Narayanan, R. Pieraccini (2002), *Combining Acoustic and Language Information for Emotion Recognition*, Proceeding of the 7th International Conference on Spoken Language Processing, pp. 873-876.

[10] C. Zhang, W. Zuo, T. Peng, and F. He (2008), *Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel*, Proceedings of the Third International Conference on Convergence and Hybrid Information Technology, Vol. 2, pp. 909-914.

[11] H. Tang, S. Tan, X. Cheng (2009), *A survey on sentiment detection of reviews*, Expert Systems with Applications, Vol. 36, No. 7, pp. 10760-10773.

[12] J. Chen, H. Huang, S. Tian, and Y. Qua (2009), *Feature selection for text classification with Naïve Bayes*, Expert Systems with Applications, Vol. 36, No. 3, pp. 5432-5435.

[13] J.C. Na, C. Khoo, and P.H.J. Wu (2005), *Use of Negation Phrases in Automatic Sentiment Classification of Product Reviews*, Library Collections, Acquisitions, and Technical Services, Vol. 29, No. 2, pp. 180-191.

[14] K. Khan, B. B.Baharudin, A. Khan, F. e-Malik (2009), *Mining Opinion from Text Documents: A Survey*, The 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 217-222.

[15] K. Polat and S. Gunes (2009), *A New Feature Selection Method on Classification of Medical Datasets: Kernel F-score Feature Selection*, Expert Systems with Applications, Vol. 36, No. 7, pp. 10367-10373.

[16] M. Simeon and R. Hilderman (2008), *Categorical Proportional Difference: A Feature Selection Method for Text Categorization*, Proceedings of the 17th Australasian Data Mining Conference, pp. 201-208.

[17] M. Karabatak and M.C. Ince (2009), *A New Feature Selection Method Based on Association Rules for Diagnosis of Erythemato-squamous Diseases*, Expert Systems with Applications, Vol. 36, No. 10, pp. 12500-12505.

[18] P. Chaovalit and L. Zhou (2005), *Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches*, Proceedings of the 38th Hawaii International Conference on System Sciences.

[19] Q. Ye, Z. Zhang, and R. Law (2009), *Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches*, Expert Systems with Applications, Vol. 36, No. 3, pp. 6527-6535.

[20] R. Prabowo and M. Thelwall (2009), *Sentiment analysis: A combined approach*, Journal of Informetrics, Vol. 3, No. 2, pp. 143-157.

[21] S. Tan and J. Zhang (2008), *An Empirical Study of Sentiment Analysis for Chinese Documents*, Expert Systems with Applications, Vol. 34, No. 4, pp. 2622-2629.

[22] T. O'Keefe and I. Koprinska (2009), *Feature Selection and Weighting Methods in Sentiment Analysis*, Proceedings of the 14th Australasian Document Computing

Symposium.

[23] T. Wang, H. Huang, S. Tian, and J. Xu (2010), *Feature Selection for SVM via Optimization of Kernel Polarization with Gaussian ARD Kernels*, Expert Systems with Applications, Vol.37, No. 9, pp. 6663-6668.

[24] V.N. Vapnik (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag.

[25] W. Zhang, T. Yoshida, X. Tang (2011), *A Comparative Study of TF-IDF, LSI and Multi-words for Text Classification*, Expert Systems with Applications, Vol. 38, No. 3, pp.2758-2765.

[26] X. Bai (2010), *Predicting consumer sentiments from online text*, Decision Support Systems, doi:10.1016/j.dss.2010.08.024.

[27] X. Tian and W. Tong (2010), *An Improvement to TF: Term Distribution Based Term Weight Algorithm*, Proceedings of the second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), pp. 252-255.

[28] Z. Zhang, Y.J. Li, Q. Ye, and R. Law (2008), *Sentiment Classification for Chinese Product Reviews Using an Unsupervised Internet-based Method*, Proceeding of the 15th International Conference on Management Science and Engineering, pp. 3-9.