

部落格探勘

Blog Mining

陳隆昇*

朝陽科技大學

資訊管理學系

lschen@cyut.edu.tw

杜逸普

朝陽科技大學

資訊管理學系

s9514629@cyut.edu.tw

摘要

繼電子郵件、電子佈告欄、即時通訊之後，部落格(Blogs)已逐漸成為一個足以改變世界的網路應用。近來，有越來越多的企業嘗試在部落格上，挖掘有利於商業目的的資訊以及知識。因此，本研究最主要的目的，是提出一個創新的部落格探勘模組(Blog Mining Model)來擷取部落格上的知識，並且能夠達到下列目的：1. 顧客區隔：我們利用自我組織特徵映射圖(Self-organizing Map, SOM)，來區隔市場。2. 顧客分類：以隱含語意索引(Latent Semantic Indexing, LSI)屬性攫取(Feature Extraction)方法為基礎，利用倒傳遞類神經網路(Back-propagation Neural Networks, BPN)以及決策樹(Decision Tree)2種分類方法，來發展處理稀疏性資料的分類器，並且預測新進客戶所屬的分類。最後，本文提供一個關於網路電話產品的實例，來驗證所提方法的有效性。

關鍵字：部落格、資料探勘、自我組織特徵映射圖、分類、隱含語意索引

Abstract

A blog is a user-generated website and becomes a new and popular media, especially for new generation of Internet users. Depending on the blog search engine, Technorati, we can track about 60 million blogs in January 2007. Therefore, blogs are viewed as new marketing channels today. Recently, more and more companies attempt to discover useful information and knowledge for business purposes. Therefore, the major objective of this study is to present a new structure for extracting knowledge from blogs, which is called Blog-Mining

Model.

This model achieved two purpose. One can make the market segmentation via Self-organizing Map (SOM), and another can predict the new customer's class via Back-Propagation Neural Networks (BPNN) and Decision Tree. Finally, we utilize Latent Semantic Indexing (LSI) to improve a sparse data classifier. A real case regarding VoIP phone will be provided to illustrate the superiority of our proposed methodology.

Keywords : Blog, Data mining, Self-organizing Map, Classification, Latent Semantic Indexing.

1. 前言

部落格 (Blog) 一詞，是由 Web log 所衍生而來，原本指的是網路伺服器所記錄的訪客記錄檔。後來 John Barger (1997) 提出「Weblog」一詞，代表使用網頁作為媒介，呈現個人記錄。最後 Peter Merholz (1999) 將 Weblog 分開唸為「We Blog」，意思是讓我們一起來寫部落格，Blog 變成了動詞，指的是一種在網路上發表感想或是撰寫日誌的行為，每個人都可以在網路上，自由的發表自己的言論，而撰寫或經營部落格的人，便稱為部落客 (Blogger) 或博客。由於部落格具備了資料的真實性、資訊的即時性，以及使用操作上的便利性，使得部落格在近幾年內快速成長，成為一種新興的傳播媒體。

微軟 (Microsoft) 董事長比爾·蓋茲 (Bill Gates) 曾說，部落格是繼電子郵件 (E-Mail)、電子佈告欄 (Bulletin Board System, BBS) 與即時訊息 (Instant Messaging, IM) 之後，第四個改變世界的網路應用[5]。在台灣，部落格的使用情形從2005年約有61.2%的網友擁有自己的部落格[20]，到2006年有8成的網友已使用

過部落格，其中閱讀他人部落格有 69%，略多於自己編輯的 55%[19]，2007 年則有 93.8% 的網友瀏覽過部落格，有 66.6% 的網友擁有自己的部落格[21]。在全球，部落格的使用數量於 2007 年 4 月也已突破 7200 萬個，是 2005 年 3 月的 9 倍[22]。從這些數據，我們可以發現，部落格在台灣已經是新興的一股科技熱潮，在全球更是不容忽視的力量。

部落格的使用，除了受到一般的網友的喜愛之外，更有許多企業利用部落格來從事商業活動，如通用汽車（General Motors）成立一個官方部落格——Fastlane，讓消費者與企業主管在部落格上交流各種汽車的心得與意見。通用汽車可以在第一時間聆聽消費者心聲，一方面降低錯誤策略所帶來的損失，另一方面助於評估未來策略發展的正確方向[4]。由此可見，部落格已從早期以網頁形式呈現個人記錄，或者是網友抒發心情的網路日誌，逐漸的受到企業的重視，成為企業宣傳、廣告及行銷的強力媒體，在資訊經濟的時代中，為企業帶來了新的商機與經營模式。而以往利用個人網頁作為宣傳、廣告及行銷工具等企業，也因為更新或維護網頁耗時費工，以及必須投入較高的資金及技術[3]，逐漸的捨個人網頁不用，而改用部落格作為行銷媒體。

總而言之，部落格不論是在個人的使用上，或者是企業的應用上，都富涵著潛在的商業資訊及知識，而透過資料探勘的技術，可以將這些潛在的商業資訊與知識擷取而出，並加以應用。基於擷取潛在於部落格文章商業資訊與知識的立場，本研究將提出一個『部落格探勘模組（Blog Mining Model）』，來針對部落格上的文章做資料探勘，並且以網路電話（Voice over Internet Protocol, VoIP）產品為研究對象進行探討，擷取出隱含於部落格文章中，關於網路電話的商業資訊以及知識。

透過本研究提出之部落格探勘模組，對關於網路電話的部落格文章進行資料探勘後，我們預期可以達到下列研究目的：

1. 顧客區隔：透過自我組織特徵映射圖（Self-organizing Map, SOM）的群聚技術，我們得以將部落格上的文章得以分為數群，然而每一篇部落格的文章代表了每一位發表心得的顧客，所以我們可以依照顧客發表的文章內容，將顧客區隔成數群的網路電話使用者，供行銷決策使用。
2. 顧客分類：我們可以從顧客的部落格文章，攫取知識，建立分類器，除了幫助企

業瞭解顧客對網路電話的需求外，更可將未知的新進顧客分類在適當的消費群，一方面協助顧客選擇適合的網路電話產品，另一方面也協助企業針對不同消費群的客戶，推銷合適的網路電話產品，以滿足顧客所需。

3. 發展處理稀疏資料（Sparse Data）分類器：在本研究提出的部落格探勘模組第二階段中，針對部落格文章的稀疏性（Sparsity）可能會造成分類績效的影響，我們導入『隱含語意索引』（Latent Semantic Indexing, LSI）方法，作為屬性擷取（feature extraction）工具，希望能有效的改善空間浪費、影響分類預測準確率等稀疏資料的缺點，進而提高分類預測的正確率。

2. 文獻探討

2.1 分群

分群，簡單來說，就是將類似的抽象物件或實體聚集，形成數個類別類別叢集的過程，即稱為分群。在同一叢集中，物件與物件彼此的相似度非常高；在不同叢集中，物件與物件彼此的相異性非常高。分群與分類（Classification）的差異，是分群要劃分的類別是未知的，而分類所要劃分的類別是已知的[6]。

在眾多分群方法中，楊東昌[7]認為 SOM 為一個有效並且方便於觀察的分群工具，特別是針對大量且高維度的資料分析，使用上也不需要先備知識。又 SOM 可針對大量高維度的資料進行視覺化，且 SOM 演算法與網路結構簡單並易於延伸應用，同時又具備資料降維與拓樸保存的特性，自 Kohonen 從 1982 年[16]提出至今，已經成功應用於不同領域中[16][17]。

SOM 是一種非監督式學習及競爭式學習的類神經網路，透過將高維度的資料，映射到一個二維的格子拓樸圖上，能夠將這些複雜的高維度資料有效的視覺化[17]。在映射的過程中，是利用輸入資料的訓練，並以資料的特徵向量進行訓練，將高維度的資料映射到二維的地圖上，以達到資料維度的縮減。此時，映射於二維拓樸圖上的資料，相似性高的資料將會鄰近在一起；而相異性高的資料則不會鄰近，甚至在二維地圖的對角上，利用 SOM 這樣的特性，可以應用於市場區隔上[12]。

SOM 演算是以特徵映射的方式，將任意維度的輸入向量，映射到較低維度的特徵映射圖上，形成拓樸層架構的映射圖，如圖 1 所示，

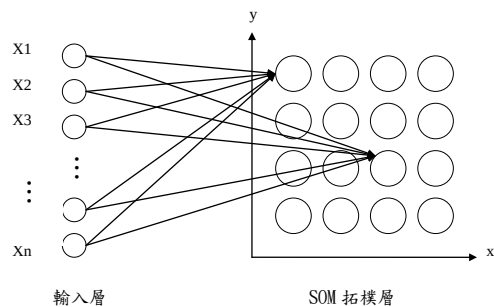


圖 1 映射至二維拓模圖之 SOM 架構圖

並且依據目前的輸入向量在神經元之間彼此相互競爭，優勝的神經元可以獲得調整連結權重向量的機會。最後輸出層的神經元會依據輸入向量的特徵，以有意義的拓模架構展現在輸出空間中，並由拓模結構圖來反應出所有輸入值之間的分佈關係。因此將此網路稱為自我組織特徵映射圖，而輸出的映射圖也稱之為拓模圖[2]。

2.2 分類

分類，是一種「先判斷明確的類別標籤(Class Labels)，再利用訓練資料(Training Data)以及在分類屬性中的類別標籤，學習出相關分類規則，之後將這些規則用來分類新進的資料」[11]。分類的模式主要有兩個步驟，第一步驟先建立描述資料的類別集合，然後透過分類演算法分析訓練資料，以產生分類相關規則，第二個步驟是利用測試資料(Test Data)來評估先前產生的分類規則之正確性，如果正確性高，則這些規則即可被應用於新進資料的分類上[11]。

2.2.1 倒傳遞類神經網路

倒傳遞類神經網路 (Back-propagation Neural Networks, BPN) 是目前類神經網路學習模式中，最具有代表性，應用也最普遍的模式[9]。BPN 是一種監督式學習網路，並具有學習 (Learning) 精度高、回想 (Recall) 速度快、輸出值可為連續值、能處理複雜的樣本識別以及高度非線性的函數合成問題，例如：樣本識別 (Sample Recognition)、分類、預測 (Prediction)、雜訊過濾 (Noise Filter)、資料壓縮 (Data Compression) 等問題[2]。BPN 主要的架構為輸入層、隱藏層及輸出層，如圖 2 所示；演算過程則是利用最陡坡降法 (The Gradient Steepest Descent Method)，將 BPN 輸出值與實際輸出值之誤差 (Error) 降低至最小

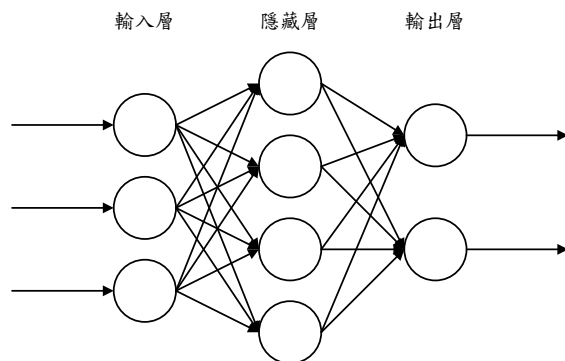


圖 2 BPN 網路架構圖

(資料來源：修改自[2])

[18]。

BPN 執行的步驟如下：

- 步驟 1 設定網路參數
- 步驟 2 亂數產生初始連結權重值及閾值
- 步驟 3 計算隱藏層與輸出層之輸出值
- 步驟 4 計算目標函數
- 步驟 5 計算權重修正量
- 步驟 6 調整各層之連結權重及閾值
- 步驟 7 判斷是否仍有訓練樣本，若是則回到步驟 3，若否則到步驟 8
- 步驟 8 判斷是否達到網路停止原則，若是則停止網路訓練，若否則回到步驟 3

2.2.2 決策樹

決策樹是一種眾所皆知的法則歸納演算法，可以用來挖掘知識，常被應用在分類或者是預測的問題上[15]，和 BPN 同屬於監督式學習。決策樹是以樹狀資料結構呈現，包含了一個樹根 (Root)、數個節點 (Node) 以及數個樹葉 (Leaf)，每個節點都代表著一個屬性 (Attribute) [13]。每一個從樹根到樹葉的分支，就代表著一個分類的規則，而決策樹的優點正是能產生容易被人類所瞭解並應用的決策法則[8]。

決策樹是利用樹狀資料結構的為基礎的分類方法，優點是能產生讓人類容易瞭解的決策規則。決策樹的建構是以監督式學習方法建立，每個內部節點 (Internal Node) 為某屬性的測試，而分支 (Branch) 代表對某屬性測試的結果，最後的樹葉節點 (Leaf Node) 代表類別。當一筆資料進入根部後，依照演算法則開始向子節點 (Child Node) 歸納，最後分類到適合的樹葉節點，亦為找到適合的類別。從根部進入，最後找到適合的類別，這樣的路徑稱為分類規則。如圖 3 為一個典型的決策樹範例。

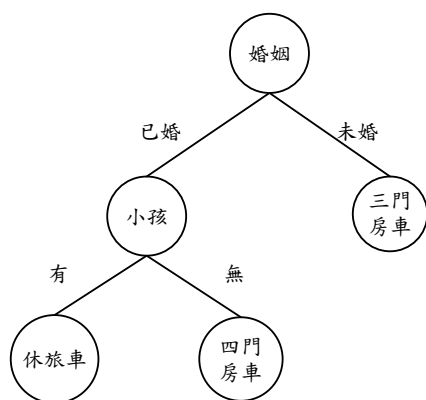


圖 3 決策樹狀圖
(資料來源：修改自[1])

2.3 隱含語意索引

由於利用詞彙索引為基礎的檢索方法，向來面對著兩個影響索引結果的問題：一字多義以及一義多字，主要是因為同樣的概念，可以用不同詞彙來表達，而這樣的問題便會影響檢索的準確性。因此 Deerwester 等人[10][14]提出 LSI 方法，來改善上述的問題。

隱含語意是指同一個概念下，語法上不同的表達方式。LSI 是建構在向量空間模型 (Vector Space Model) 的基礎上，並以奇異值分解 (Singular Value Decomposition, SVD) 的方法，將原本的詞彙和文章以簡化的 k 個概念 (Concept) 表示，縮小空間維度，將相關的詞彙及文章聚集在一起，以解決資訊檢索中，隱含語意的問題[6][10]。

在此，本研究將 LSI 視為一種屬性擷取工具，希望能藉由 LSI 能解決在資訊檢索上的困擾，以及向量空間維度縮減的特性，發展改善稀疏性資料分類器。我們分別採用 Global-LSI 如圖 4 所示，以及 Local-LSI 如圖 5 所示，兩種方法進行。Global-LSI 是不管類別分佈的結果，直接針對屬性進行維度縮減，而 Local-LSI，是分別依照每個不同的類別，進行維度縮減，換言之，與 Global-LSI 相較，Local-LSI 多考慮了類別的資訊。接著取 k 維度

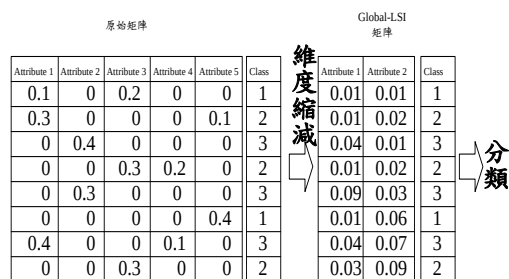


圖 4 Global-LSI 示意圖



圖 5 Local-LSI 示意圖

後，以維度縮減後的 A_k 向量矩陣進行分類預測的實驗，並與未施行 LSI 方法比較，最後驗證是否有效改善稀疏性資料的分類準確率。

3. 部落格探勘模組

本研究主要目的，是提出一個「部落格探勘模組」，並利用這個模組來擷取出隱含於部落格文章中，富涵商業資訊之知識。我們期望可以透過擷取出的知識，達到：1. 顧客區隔，2. 顧客分類，以供行銷決策使用如圖 6 之第一、二階段。同時我們也發展處理稀疏性資料分類器，提高分類預測的正確率。在本節中，我們將詳細的介紹本文提出的「部落格探勘模組」所使用的相關研究方法。其研究流程如圖 6，詳細步驟說明如下：

- 步驟 1 部落格文章蒐集：針對欲探討的主題，進行部落格上的文章蒐集，在本論文中，我們以網路電話產品的文章為探討的主題，因此對部落格上討論網路電話產品的文章進行蒐集。
- 步驟 2 資料前處理：任何欲進行資料探勘的資料，都必須先經過資料前處理，才容易挖掘出有用的資訊及知識。本論文中，

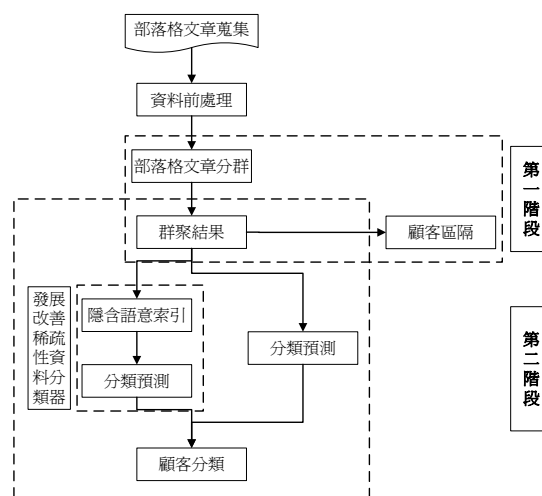


圖 6 部落格探勘模組流程圖

- 我們是利用詞彙-文章向量矩陣作為訓練輸入資料，因此我們資料前處理的流程是定義關鍵字、刪去贅詞關鍵字、合併關鍵字、刪去未出現關鍵字以及產生詞彙-文章向量矩陣，並對詞彙-文章向量矩陣進行正規化。
- 步驟 3 資料分群：區隔這些討論網路電話產品的部落客，觀察不同群聚裡的部落客有不同的需求面，希望藉此達到顧客區隔。因此在資料分群的技術上，我們使用的是自我組織特徵映射圖，利用部落客所發表的文章，依照關鍵字特色，能夠自我形成群聚，實現顧客區隔的目的。
- 步驟 4 資料分類：透過既已蒐集網路電話產品文章的顧客區隔結果，來預測未來發表網路電話產品文章的部落客會是屬於既有群聚中的哪一群，以供決策行銷之用。因此我們使用倒傳遞類神經網路、支撐向量機及決策樹等監督式學習的分類預測技術，來進行資料分類預測。
- 步驟 5 隱含語意索引：由於詞彙-文章向量矩陣屬於稀疏性資料的矩陣，往往耗費儲存空間以及影響分類預測準確率，因此本論文希望能藉由導入隱含語意索引技術，發展能有效改善稀疏性資料分類器，提高分類預測的準確率。

4. 實驗結果

本章節實際蒐集發表於部落格之網路電話文章，實證本研究提出之部落格探勘模組。實驗資料敘述如下：

以 Yahoo!奇摩部落格、無名小站部落格為主，蒐集兩百筆討論網路電話的文章。文章範例如：

(1) 我下午掛完跟你們的網路電話，馬上跟我同學連線，我跟他的通話就非常非常清楚，跟在台灣講電話一樣，完全不會 LAG 應該要這樣才對，不知道是不是我們家寬頻太慢，我搞清楚再跟你們說，要是是寬頻問題，你們就去升等好了，你們今天跟我對話的軟體叫做 skype，說明書在電腦桌子墊子下面，視訊我今天是跟你用 MSN，所以有兩種軟體，兩種都可以語音跟視訊，不用錢請多利用。

(2) 今天用 skype 的網路電話，通話費是很便宜，但是通話品質不是很好，還有很大的進步

空間，剛開始不會斷訊，過了兩三分鐘後就開始會斷訊了，不好的地方在這，傳話的速度也是有點太慢了。

4.1 資料前處理

資料蒐集完畢後，接著即是進行資料前處理：

- 步驟 1 定義關鍵字：以「網路電話服務品質」、「網路電話產品品質」以及「火星文集使用者情緒字眼」構面來定義關鍵字，共定義出 55 個關鍵字，如表 1 所示。
- 步驟 2 刪去贅詞之關鍵字：將不具意義、口吻性用語之關鍵字刪去，如表 2 所示。
- 步驟 3 整合關鍵字：將相似意義之關鍵字合併，如表 3 所示。
- 步驟 4 刪去未出現或出現頻率極低之關鍵字：未出現及出現頻率低，在本研究不具任何意義，將其刪去，如表 4 所示。
- 步驟 5 產生詞彙-文章向量矩陣：經過了步驟 1 至步驟 4 的階段，我們可以得到最後的關鍵字定義表，如表 5。接著我們將部落格上所蒐集的 200 筆網路電話文章，依照表 5 所列的 33 個關

表 1 關鍵字定義表

網路電話服務品質構面			
通話監聽	通話干擾	通話延遲	通話品質
增值服務	通話費率	通話費	配號
帳單	多方會談	國際電話	國內長途
頻寬	斷線率	網內互打	Skype
MSN	撥通成功率		
網路電話產品品質構面			
麥克風	藍芽	免持聽筒	擴音
重量	喇叭	迴音	液晶螢幕
視訊	來電鈴聲	來電保留	待機時間
通話時間	電池	網路攝影機	
火星文及消費者情緒字眼構面			
Orz	囧	^_^	XD
>_<	貴	:)	便宜
:(	= =	方便	清晰
模糊	省錢	免費	好
不好	會	不會	有
沒有	讚		

表 2 刪去贅詞關鍵字

好	不好	有	沒有	會	不會
---	----	---	----	---	----

表 3 整合前後之關鍵字

整合前關鍵字	整合後關鍵字
通話干擾 通話延遲 通話品質	通話品質
通話費率 通話費 帳單	通話費用
斷線率 撥通成功率	斷線率
藍芽 免持聽筒	無線傳輸
待機時間 通話時間 電池	電池
Orz 囧	Orz
^_ :))	:)
便宜 省錢	省錢
>_ :(= =	: (

表 4 刪去未出現之關鍵字

來電保留	液晶螢幕	重量
------	------	----

表 5 前處理後之關鍵字定義表

通話監聽	通話品質	加值服務	通話費用
配號	多方會談	國際電話	國內長途
頻寬	斷線率	網內戶打	SKYPE
MSN	麥克風	無線傳輸	擴音
喇叭	迴音	視訊	來電鈴聲
電池	Orz	:)	XD
: (	貴	省錢	方便
清晰	模糊	免費	讚
網路攝影機			

鍵字，產生一個詞彙為 33、文章為 200 的詞彙-文章向量矩陣。

步驟 6 向量矩陣正規化：將步驟 5 所得之向量矩陣值進行正規化，使得矩陣內的向量值介於 0~1 之間，方便我們在往後的研究步驟中進行實驗。

4.2 資料分群

以 Matlab 7.0.1 執行 SOM，初步實驗參數設置為：輸入資料維度為 33 個關鍵字詞彙*200 筆文章，輸出拓樸層為 1*3 的二維地圖，粗調學習速率為 0.9，細調學習速率為 0.1，半徑距離為 1，終止條件為迭代次數 100 次。執行完畢後，得到群聚如表 6~表 8 所示。

從表 6 來看，可以發現第一類除了基本的通話費用、品質外，他們還著重在「麥克風」及「視訊」上。這說明著，會在部落格發表這些文章的使用者，他們希望的是功能多樣化的網路電話。因此，本文將第一類的用戶定義為「玩家級」的網路電話部落客。

在表 7，第二類有相當高的比例是在探討著「通話品質」。那麼發表這些注重網路電話通話品質文章的部落客，可能是企業相關的用者，因為對於企業而言，每一通商業電話，都可能是

表 6 第一類前 70%累加比例表

第一類	關鍵字	比例	累加比例
	Skype	19.08%	19.08%
	麥克風	9.92%	29.01%
	通話費用	9.16%	38.17%
	視訊	8.65%	46.82%
	省錢	7.89%	54.71%
	通話品質	5.85%	60.56%
	國際電話	5.09%	65.65%
	MSN	4.07%	69.72%

表 7 第二類前 70%累加比例表

第二類	關鍵字	比例	累加比例
	Skype	27.69%	27.69%
	通話品質	20.92%	48.62%
	省錢	6.46%	55.08%
	通話費用	6.15%	61.23%
	MSN	5.54%	66.77%
	麥克風	4.92%	71.69%

表 8 第三類前 70%累加比例表

第三類	關鍵字	比例	累加比例
	Skype	18.92%	18.92%
	免費	11.64%	30.56%
	省錢	11.02%	41.58%
	通話費用	9.98%	51.56%
	通話品質	8.73%	60.29%
	國際電話	7.28%	67.57%
	網內互打	4.57%	72.14%

一筆重要的交易，因此他們會非常要求網路電話的品質，其次才是通話費用。所以，本文將第二類定義為「企業用戶」的網路電話部落客。

從表 8 來看，可以發現第三類有百分之五十的比例，著重在「免費」、「省錢」、「通話費用」等項。這意味著，這一類的文章，是非常重視網路電話的通話金額，在部落格上發表這樣的文章，可能是一些對網路電話要求不高的，只希望有便宜，甚至是免費的網路電話可以使用的部落客所發表的。本文將第三類定義為「經濟實惠」的網路電話部落客。

因此，透過資料分群的方法，本文將發表網路電話產品文章之部落客，分為三類：

第一類：玩家級網路電話部落客。

第二類：企業用戶網路電話部落客。

第三類：經濟實惠網路電話部落客。

4.3 資料分類

在這本節中，我們將利用群聚結果所得到的顧客類別，作為我們分類預測的類別標籤，並使用 BPN 以及決策樹進行分類預測，以預測未來新進顧客的分類準確率。

本研究以 Matlab 7.0.1 來執行 BPN，是一個輸入層神經元 33 個，隱藏層神經元 17 個，輸出層神經元 1 個的架構，學習速率為 0.3，動量係數為 0.95，迭代次數為 1000 次。決策樹則是以 See 5 來執行。我們取 80%比例的文章作為訓練資料，剩下的 20%比例的文章作為測試資料，並且將資料切成 5 個等分資料，以作為交互驗證。分類預測結果如表 9 所示。

4.4 以 LSI 為基礎的資料分類器

在本節，我們嘗試發展改善稀疏性資料分類器，因此我們導入 LSI 技巧，觀察其分類預測的結果，並且和未導入 LSI 的分類預測結果比較。BPN 的輸入層神經元為 LSI 縮減的 k 維度個，隱藏層神經元為(輸入層+輸出層)/2 個，輸出層為 1 個，學習速率為 0.1~0.3，動量係數為 0.7~0.95，迭代次數為 1000 次。決策樹以

表 9 資料分類預測結果

	BPN	決策樹
Fold 1	82.50%	90.00%
Fold 2	85.00%	92.75%
Fold 3	92.50%	77.50%
Fold 4	92.50%	77.50%
Fold 5	85.00%	75.00%
平均正確率	87.50%	82.55%

表 10 Global-LSI+ 資料分類預測結果

	Global-LSI +BPN	Global-LSI +決策樹
Fold 1	92.50%	90.00%
Fold 2	97.50%	97.50%
Fold 3	95.00%	95.00%
Fold 4	95.00%	87.50%
Fold 5	87.50%	80.00%
平均正確率	93.50%	90.00%

表 11 Local-LSI+ 資料分類預測結果

	Local-LSI +BPN	Local-LSI +決策樹
Fold 1	95.00%	100.00%
Fold 2	80.00%	82.50%
Fold 3	95.00%	97.50%
Fold 4	95.00%	100.00%
Fold 5	97.50%	97.50%
平均正確率	92.50%	95.50%

See 5 執行，LSI 則為 Global-LSI 與 Local-LSI 兩種方法執行。Global-LSI 的結果如表 10 所示，Local-LSI 的結果如表 11 所示。

5. 結論

對網際網路使用者而言，部落格已經成為不可或缺的工具，每位使用者可以藉由部落格抒發個人的心情或想法，也因此，有越來越多企業希望在部落格挖掘到有用的知識以供決策行銷之用。因此，本研究之目的，是提出一個能從部落格上擷取知識的部落格探勘模組。

透過本研究提出之部落格探勘模組，我們能夠做到：

1. 顧客區隔，將使用網路電話的部落客區分為玩家級部落客、企業用戶部落客以及經濟實惠部落客等三類。
2. 建立一個預測的分類器，來預測新進顧客所屬的類別。
3. 利用 LSI 改善分類器對於稀疏性資料預測分類的準確性。

參考文獻

- [1] 林盈源，決策樹在資料庫行銷決策之應用，碩士論文，國立成功大學工業管理科學系專班，台南，2003。

- [2] 張斐章、張麗秋、黃浩倫，類神經網路：理論與實務，台灣東華書局股份有限公司，2003。
- [3] 戈立秀，部落格之資訊蒐集與分享行為之研究，碩士論文，國立台灣大學圖書資訊學研究所，台北，2007。
- [4] 陳品均，Web2.0 應用服務策略行動之研究—以 Yahoo!、Google、MSN 為例，碩士論文，國立台灣大學商學研究所，台北，2006。
- [5] 陳穆臻，Blog 行銷有多行?!，*管理雜誌*，第 376 期，第 114-117 頁，2005。
- [6] 曾韋榮，結合潛在語意檢索及資訊粒化於資料探勘，碩士論文，國立台北科技大學商業自動化與管理研究所，台北，2006。
- [7] 楊東昌，自組織映射圖神經網路改善模式與分群應用之回顧研究，碩士論文，華梵大學工業管理學系，台北，2004。
- [8] 尤春惠，資料探勘在用藥安全上的應用：預測泛可黴素在腎衰竭病患上的用量適當性，碩士論文，國立中山大學資訊管理學系，高雄，2004。
- [9] 葉怡成，類神經網路模式應用與實作，第七版，儒林圖書有限公司，2002。
- [10] 吳忻萍，以隱藏語意索引為基礎之中文資訊檢索，碩士論文，國立台灣大學資訊管理學系，台北，1997。
- [11] J. Han and M. Kamber, Data Mining : Concepts and Techniques, *Mogan Kaufmann Publishers*, 2001.
- [12] M. Y. Kiang, M. Y. Hu, and D. M. Fisher, "An Extended Self-organizing Map Network for Market Segmentation — A Telecommunication Example," *Decision Support Systems*, Vol. 42, Issue 1, pp. 36-47, 2006.
- [13] R. A. Kumar, V. Sugumaran, B. H. L. Gowda, and C. H. Sohn, "Decision Tree- A Very Useful Tool in Analysing Flow-induced Vibration Data," *Mechanical Systems and Signal Processing*, Vol. 22, Issue 1, pp. 202-216, 2008.
- [14] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, Vol. 41, Issue 6, pp. 391-407, 1990.
- [15] S. Lee, S. Lee, and Y. Park, "A Prediction Model for Success of Services in E-commerce Using Decision Tree—E-customer's Attitude Towards Online Service," *Expert Systems with Applications*, Vol. 33, Issue 3, pp. 572-581, 2007.
- [16] T. Kohonen, "Self-organizing Formation of Topologically Correct Feature Maps," *Biological Cybernetics*, Vol. 43, No. 1, pp. 59-69, 1982.
- [17] T. Kohonen, "The Self-organizing Map," *Neurocomputing*, Vol. 21, No. 1, pp. 1-6, 1998.
- [18] T. L. Lee, "Back-propagation Neural Network for The Prediction of the Short-term Storm Surge in Taichung Harbor, Taiwan," *Engineering Applications of Artificial Intelligence*, Vol. 21, Issue 1, pp. 63-72, 2008.
- [19] 資策會資訊市場情報中心：
<http://mic.iii.org.tw/intelligence/>
- [20] 2005 年蕃薯藤網路調查報告：
<http://survey.yam.com/survey2005/index2.html>
- [21] 波特仕線上市調網：
<http://www.pollster.com.tw/report/20070720/index.htm>
- [22] Blog 搜尋引擎：
<http://technorati.com>