

應用文件分群與文字探勘技術於機器學習領域趨勢分析以SSCI資料庫為例

Trend Analysis in Machine Learning Research from SSCI Database by Document Clustering Manipulation and Text Mining Methodology

尹其言* (Yin, Chi-Yen)、楊建民** (Yang, Jiann-Min)

(Received August 15, 2009; Revised First November 13, 2009; Revised Second November 4, 2010; Accepted November 11, 2010)

摘要

機器學習領域期刊文獻的研究與發表，一直是電腦科學未來應用與新科技誕生的基礎，本研究利用SSCI資料庫中與機器學習應用相關研究文獻，使用文字探勘技術，擷取具文章鑑別力之特徵詞彙，進行詞彙叢聚分析，將每份文章出現各詞彙叢聚的頻率做為自組織映射網路的輸入變數，利用網路神經元自動群集的功能，將機器學習應用的分成10大領域，最後配合文章發表年份進行趨勢分析，找出各研究領域的歷史脈絡，並進一步預測未來可能趨勢。

關鍵詞：文件分群、文字探勘、自組織映射網路

Abstract

This paper introduces the new concept for data mining manipulation. The research utilizes a document clustering technology to gain the homogeneous glossaries in each article at SSCI database, and forwarding toward onto the literature cluster assay. To select the term frequency indexes which generated by the glossaries aggregation as the parameters of the Self-Organization Map (SOM) network, proceeding the network neuron automatic clustering function, it is to strengthen the discovering ability through the historical tracking and gathering the results from various research domain, and forecasting the future possible research tendency.

Keywords: Document Clustering, Text Mining, Self-Organization Map

* 國立政治大學資訊管理學系博士候選人

Doctoral Student, Department of Management Information Systems, National Chengchi University

** 國立政治大學資訊管理學系教授，國立政治大學商學院經營管理碩士學程(EMBA)執行長

Professor, Department of Management Information Systems; CEO, EMBA program, College of Commerce, National Chengchi University

壹、緒論

資訊技術快速的發展，環保節約意識興起。傳統紙本文件已逐漸為電子化形式所取代，電子化文件之傳遞可打破時間、空間與數量上的限制，譬如各類的電子期刊，可提供授權的研究人員隨時下載運用，惟電子化文件也因成長快速，致使用者有資訊過量之憾，若無適切的分類方法，不僅耗時、費力，亦無法進行有效管理與利用。文件自動分類可以用機器學習的方式，只考慮文件與文件之間的特性，對文件加以有效管理，不僅具有效率，其處理的結果亦屬客觀且一致。網際網路興盛後資訊爆炸性的成長，欲由大量資料中萃取出有用的資訊，文字探勘 (text mining) 技術因應而生。文字探勘結合資料採掘 (data mining) 技術與自然語言處理、資訊檢索技術，使大量的文字資訊能經由電腦分析歸納，主要應用包括自動分類、自動摘要、文件檢索、知識管理及趨勢預測等。

目前在文字探勘領域中，文件分類的相關研究十分廣泛，包括企業內部的公文、個人電子郵件的管理、新聞的分門別類、網頁的分類，都是和文件分類 (document clustering) 的議題相關。Lam, Ruiz, and Srinivasan (1999) 指出將文件自動分類應用在資訊檢索的研究，可改善檢索文件的品質；Moens and Dumortier (2000) 則研究將分類應用在推荐期刊雜誌給有興趣的讀者，節省讀者找尋資料與閱讀摘要的時間；而Sebastiani (2002) 的研究指出，數位文件與日俱增、大量的文件等待分類且需分類的反應時間極短、自動分類可以增加人工分類的生產率、自動分類技術漸趨成熟可媲美人工專家分類結果等因素，使得文件自動分類變得日益重要。因此，文字探勘可有效地支援文件分類與管理，並且在檢索文件時，能夠正確獲得相關的文件內容，提高檢索的精確率。

貳、文獻探討與相關研究

本研究以機器學習相關期刊為研究主體，結合特徵詞 (homogeneous glossary) 擷取、文字探勘及詞彙叢集 (glossary aggregation) 技術，針對機器學習相關期刊執行趨勢分析，相關技術將於以下各節說明之。

一、文字探勘

文字探勘係針對文字資料進行處理，透過各種量化技巧，如統計計量與資訊理論等，輔以人工智慧理論，試圖找出隱含、有趣且有助於決策之樣式。其應用領域相當廣泛，如知識管理 (Sebastiani, 2002, 2005)、資訊安全 (Garcia, Pikatza, Florez, & Sobrado, 2005)、資訊檢索 (Lu, Chien, & Lee, 2002)、自然語言處理 (Kao & Poteet, 2006) 與語意網路 (Stumme, Hotho, & Beremdt, 2002) 等。文字探勘可從大量的文件中，分析大量文字型態的資料，將文件內所隱含的資訊、知識以及這些資訊、知識的關聯性找尋出來，藉此過程來獲得文件中的資訊及知識，其主要處理的對象為電子化文件資料。

文字探勘與資料採掘的相同點在於兩者核心概念皆為找出隱含有用的樣式與知識。而相異點在於接受處理資料的結構，資料採掘也可以處理數值以

外的資料。文字不像數值具有單位統一性質，使用上較自由不受限制，因此文字探勘必須面對幾點挑戰，首先最重要的一點是文字特性的量化，藉由量化後的特性找出各文件之間的相關性，然文字文件結構通常屬於非結構或半結構化，如何訂定結構性量化的指標為文件探勘最首要的考量。其次為文件撰寫通常跟作者本身習慣與背景有相當程度的關聯性，即便針對相同事件也會有不同的敘述，如何將這些差異性降低是文件探勘必須考量的過程。第三點是文字資料亦受限於語言限制，然文件並不一定侷限於以同一種語言撰寫，文字間可能夾雜常見但屬於不同語言的詞彙，例如中文文件常包括英文字詞彙。如何將文字依據語言、語意與語法準確地辨識就扮演相當重要的環節。

現行文字探勘技術其中一種是處理分類議題，透過分類技術將大量文件區分成許多小群組，以便使用者檢索之需求。一般而言文字探勘分類技術可劃分成下列兩種形式，分別為「群集」(clustering)與「分類」(categorization)，兩者特性與意義不同，適用情境亦有所差異：

(一) 群集

群集法將文件集合切割成不同的小群集，透過切割完成後的這些小群集找出屬於該群集的主題與特性。主要原則為相同群集內文件必須具有高內聚力，群集與群集之間則必須維持低耦合力(coupling)，藉此有效的區隔群集。分群法試圖找出各群集的「樣本」(pattern)，目前樣本辨認較為常見的演算法有k平均法(k-means)、最小生成樹(minimal spanning Tree)(Moretti, 2006)、k最接近鄰居法(k nearest neighbor, k-NN)、基因演算法(Goldberg, 1989; Davis & Mitchell, 1992; Maulik & Bandyopadhyay, 2000)與階層分群(Heller & Ghahramani, 2005)等。

(二) 分類

分類法與群集法不同的一點在於分類法進行分類時必須給定事先定義好的類別集合。透過各類別中的文件子集合，辨別屬於該類別的樣式與特性。由於分類法中類別集合為事先定義，而群聚法則是依群聚特性演變自動產生，因此分類法在類別擴展上需要人工或搭配群聚法。另一方面因為類別集合為事先定義，分類法可透過訓練資料強化類別樣式之辨識力，藉此提升分類結果之準確率。

有關文字探勘的架構是由兩部分所組成，包括文字精煉(text refining)及知識淨化(knowledge distillation)：

1. 文字精煉：將不規則的文件轉換成事先已決定好的中間型式(intermediate form)。
2. 知識淨化：主要是由中間型式來推論(deduce)來整理出樣本或知識。其架構如圖1：

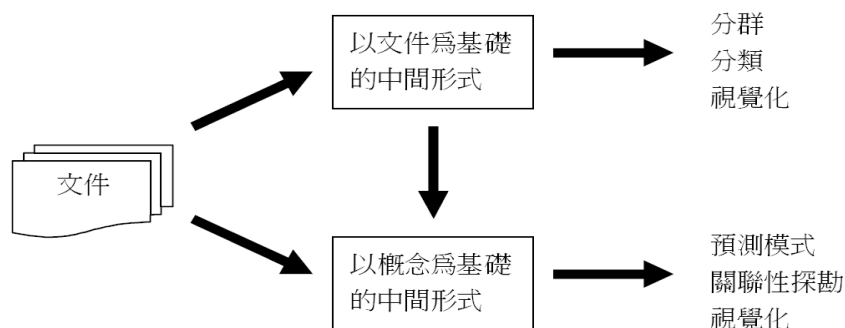


圖1 文字探勘架構圖

二、特徵詞擷取與詞彙叢聚

特徵詞彙的篩選對文件分類自動化是十分重要的前置步驟，這方面相關的研究如針對資訊增益量 (information gain)、卡方檢定 (χ^2 -statistics)、交互資訊 (mutual information)、詞彙強度 (term strength) 及文件頻率 (document frequency) 等五種特徵選取方法作比較，利用最近相鄰法及線性最小平方 (linear least squares) 等分類演算法對新聞文件進行分類；結果顯示以資訊增益量、卡方檢定與文件頻率等方法所擷取出的特徵詞彙，能提供較好的分類品質 (Yang & Pedersen, 1997)。

學者Baker與McCallum利用詞彙在各類別出現之相對機率，凝聚分布機率相仿的詞彙，並合併至一個群體內，同時也比較數個降低維度的方法，如潛在語意分析及馬可夫覆蓋特徵選取 (Markov blanket-based feature selection)、交互資訊。經過實驗證明，以叢集表示法較詞彙表示法出色，尤其是對大型測試資料集 (Baker & McCallum, 1998)。Bekkerman, El-Yaniv, Winter, and Tishby (2001) 等學者亦以類似的作法叢聚詞彙，並且以詞彙叢集對訓練文件分群，實驗結果也和Baker與McCallum等學者之結果一致。而Bassiou, Kotropoulos, & Pitas (2001) 等學者試著以詞彙分群來改善文件檢索的效果，採用階層式叢聚法對詞彙分群，並認為文件長度以及字詞在文件內出現的次數，會影響分群品質，故以傳統TF-IDF方法分配字詞權重之外，加入這兩個參數，使檢索效果更為提高。

在特徵詞數量的選擇上，很難定義重要特徵詞數量，因此本研究加入相對比較過濾雜訊的方法，利用相對性比較來過濾類別與類別之間相對不重要的特徵詞，其目的在於突顯特徵詞在類別之間的鑑別度及降低雜訊影響，以提高分類結果成效。

三、文件分類與叢聚

在文件分類的議題中，特徵詞的選擇在建立文件類別時具有重要性影響，當文件類別模型被建構時，需要對這些類別規則加以學習，以達到自動分類，其中特徵詞數量的選擇對文件分類成效上有所影響，當特徵詞到達一

定數量時，即無法再提高分類成效；甚至，會產生雜訊（noise），反而使分類成效降低。在研究的過程中發現，每一文件分類達到所屬類別時，並不是根據建構類別所使用的特徵字數量，即能夠產生最好的分類成效；某些特徵詞雖然在該文件中具有權重，但卻成為分類上的雜訊，造成分類效果不彰。

以文件本身的特性來看，文件類別的定義，很難只將文件定義到單一類別，由於不同立場、不同角度與觀點，或是人為主觀判斷，對文件類別定義來說，都有不一樣的結果，因此文件很可能被定義至不同的類別。且如果文件只被定義在某個類別，而忽略其他類別，往往會侷限了文件本身特性，並無法將文件所表達的內容、意涵充分發揮。譬如，以新聞文件來說，同樣一篇新聞的報導假使被定義在「運動」的類別中，也有可能被定義在「休閒娛樂」的類別中，因此，本研究認為文件存在眾多類別（multiple categories）特性。

文件分類可以矩陣A表示，令D為文件集合，C為類別集合，則目標方程式(1)如下：

$$\Phi : D \times C \rightarrow \{T, F\} \dots\dots\dots(1)$$

一般將 $\Phi : D \times C \rightarrow \{T, F\}$ 稱之為分類器，其中 $C = \{c_1, c_2, \dots, c_n\}$ 表示類別集合，而D表有限或無限的文件集合。當目標方程式 $\Phi(d_j, c_i) = T$ 時，表示文件 d_j 屬於 c_i 類別。文件分類中，類別僅為符號象徵，不具有任何與分類相關之資訊。常見的分類有「單一標記」與「多標記」分類、「硬式」與「排序」分類以及「文件主軸」與「類別主軸」三種（Sebastiani, 2002, 2005）。單一標記與多標記之差異在於前者進行分類時，僅有唯一一個類別被指定到每份文件，後者則可被指定任何數量之類別。分類器使用可分為以文件主軸或類別主軸。文件主軸係指將所有文件 $d_j \in D$ 給予類別標籤 $c_i \in C$ ，亦即針對每一份文件至少必須指定一個類別標籤。類別主軸導向則是將類別 $c_i \in C$ 為主體，判斷文件 $d_j \in D$ 是否歸屬於該類別，而 $d_j \in D$ 可被指定至一個以上之類別，但本質上均屬於「多標記」分類。「硬式」與「排序」分類是只完全自動分類之應用必須給定文件與類別絕對之關係，即 $\Phi : D \times C \rightarrow \{T, F\}$ ，但若針對部分自動分類應用則給定文件與類別之間一個相對關係，亦即排序值，藉由此排序值可供決策者判斷哪些類別標籤適合指定給該文件。文件特徵（feature）為文件之代表，文件特徵建構的常見方法為關鍵字選取。藉由數個關鍵字代表文件，使用者可比對關鍵字檢索文件。特徵建構將文件以向量形式表示，透過特徵與文件對應關係可得一個矩陣A，矩陣元素 a_{ik} 則表示該特徵在該文件中出現之權重。

$$A = (a_{ik}) \dots\dots\dots(2)$$

權重量化常見方式為詞彙頻率（term frequency），即 a_{ik} 值為詞彙在文件

中出現頻率。TF-IDF則將逆向文件頻率 (inverse document frequency) 納入考量 (Satlon & Buckley, 1988)。亦有文獻以文件內容長度為量化法。資訊熵 (information entropy) 函數以資訊理論為基礎發展而來，相較於前述方法較為精密複雜，但其效果也相對較佳。維度縮減為特徵建構相當重要的議題，基於效能與穩定性之考量，進行適度的維度縮減可增進分類效率 (Aas & Eikvil, 1999)。維度縮減主要以特徵選取 (feature selection) 為主。下列舉出三種常見特徵選取法，分別是「文件頻率門檻」、「資訊獲利」與「卡方統計」。以下將針對此三種方法進行敘述。文件頻率門檻 (document frequency threshold; DF) 計算所有樣本文件中特徵 (characteristic) 出現之頻率，並將低於門檻值之特徵從特徵空間中排除 (Yang & Pedersen, 1997)。其假設前提為罕見詞彙可能代表該詞彙較不具資訊意涵或對整體效能影響力過小，對分類不具貢獻度。目前發展之特徵選取方法中，以此法最為簡單有效率，對於大量樣本及分類集合具有較佳之效能，但此法通常是伴隨其他特徵選取方法一起使用。

資訊含量 (information gain; IG) 由Quinlan於1979年提出，藉由測量樣本特徵在文件中出現與否計算其資訊位元數值並用以預測分類 (Yang & Pedersen, 1997; Aas & Eikvil, 1999)。令 $\{c_i\}_{i=1}^k$ 表可能之類別集合，則詞彙 w 之資訊獲利值以下式(3)表示：

$$IG(w) = -\sum_{i=1}^k P(c_i) \log P(c_i) + P(t) \sum_{i=1}^k P(c_i / w) \log P(c_i / w) + P(\bar{w}) \sum_{i=1}^k P(c_i / \bar{w}) \log P(c_i / \bar{w}) \cdots \dots \dots (3)$$

其中 $P(c_j)$ 表類別 c_j 在訓練樣本中出現之機率。 $P(w)$ 則表詞彙 w 在訓練樣本中出現之機率。 $P(c_j / w)$ 表示當詞彙 w 出現時，該文件屬於 c_j 之機率，而 $P(c_j / \bar{w})$ 則表示當詞彙 w 沒出現時，該文件屬於 c_j 之機率。藉由此法計算每一個詞彙之資訊獲利值，並將低於門檻值之詞彙從特徵集合中刪除。文件分群是將整個文件集自動產生數個不同的群體，以便於管理及利用。文件分群的類別不需要事先定義，可由文件與文件之間用其本身的相似度將文件分群 (clustering)，可分為「階層式」和「非階層式」。

(一) 階層式分群法 (hierarchical clustering)

階層式演算法所產生的群集是一個樹狀結構、具有階層關係的群集。可依照演算法行進方向分為兩種：一種是由大到小的群集，稱為分裂式群集法；另一種是由小到大的群集，稱為凝聚式分群法 (Bassiou, Kotropoulos, & Pitas, 2001)。

1. 凝聚式 (agglomerative) 的做法是由每個資料視為各個不同的小群集，慢慢地結合形成較大群集，最後全部的資料形成一個單一群集。
2. 分裂式 (divisive) 的做法則是先將所有的資料視為一個單一群集，

再慢慢分出較小的群集，直到最後所有的資料點各別形成一個群集。

(二) 非階層式分群法 (non-hierarchical clustering)

非階層式分群法則以k-means演算法代表，k-means是一個簡單快速的分群方法，但需要事先決定分群數目，其演算法如下：

1. 選擇一個k值，k代表決定總分群數目。
2. 隨機選擇k個資料點當作最初群集中心。
3. 計算其他成員與k個類別的距離，並分配該成員到距離最接近的群集中心。
4. 當所有成員都分配到所屬類別後，重新計算各個類別新的中心距離。
5. 假如新的平均數值和先前的平均數值相同，則終止此程序。否則，使用新的平均值當作群集的中心，並且重複步驟3至步驟5。

參、實驗流程與方法

一、資料收集與前處理

本研究方法使用SSCI資料庫進行文獻收集，選取條件為文章的主題含關鍵字“machine learning”，此處所指主題含摘要、關鍵字、應用方向等SSCI電子資料庫資料欄位，發表時間為自1956年到2008年且文章類型為“article”之相關期刊文獻共554篇。由於原始資料格式紊亂，而本研究使用文字探勘軟體QDA Miner V3.1進行分析，因此需先對來源資料進行資料前處理，過濾多餘欄位（如：作者等）、刪除品質不佳資料（如：缺少摘要、資訊不足）與並轉換格式成可供詞彙分析之資料格式，如下表1所示：

表1 資料格式

	摘要	關鍵字	論文名稱	研究方向
Case1	(Text)	(Text)	(Text)	(Text)
Case2	(Text)	(Text)	(Text)	(Text)
Case3	(Text)	(Text)	(Text)	(Text)
Case4	(Text)	(Text)	(Text)	(Text)
Case5	(Text)	(Text)	(Text)	(Text)
...	(Text)	(Text)	(Text)	(Text)
Case554	(Text)	(Text)	(Text)	(Text)

肆、實驗結果與詞彙叢聚分析

一、斷字與斷詞分析

原始資料為英文字，因此需先針對原文進行斷字斷詞分析，本研究使用QDA Miner V3.1進行斷字分析，將本文擷取成個別英文字，斷字分析結果如

表2所示，資料共含96,935個英文字，其中共含各種不同詞性的詞彙共5,720個，又其中連接詞（如：and、or、as）、冠詞（如：a、the）、介系詞（如：at、above）、形容詞（如：some、specific）共383個詞彙對資料分析過程不具意義，因此將其排除在外，總計使用5,337個詞數進行特徵詞擷取。

表 2 資料語料設置表

資料名稱	分析詞彙數	詞彙總數	單字總數	文件數
SSCI 文獻電子資料庫	5,337	5,720	96,935	554

（一）特徵詞擷取處理

為能選取最具鑑別力的特徵詞彙，避免過多特徵詞彙導致詞彙群集命名困難，以往學者認為使用詞彙TF-IDF權重為前百分之十之詞彙數進行叢聚效果較好（Yang & Pedersen, 1997），本研究透過詞彙頻率門檻，選擇滿足以下三項條件：(1)選取詞彙出現頻率大於15個Case，如果詞彙出現頻率過低，不具鑑別力。(2)詞彙出現不得超過涵蓋70%以上的Case，因為詞彙在每一個Case內都出現，亦不具鑑別力。(3)結合TF-IDF方法，而詞彙 t_i 的權重表示法如下式(4)所示：

$$t_i = (w_{i0}, w_{i1}, \dots, w_{ij}, \dots, w_{iN-1}, p(w | C_0), \dots, p(w | C_{p-1})) \dots\dots\dots(4)$$

依據以上三項條件，最後僅採用TF-IDF值大於26之詞彙。

透過頻率門檻及TF-IDF權重分析，結果共篩選出533個特徵詞，占全部詞彙9.98%，用以進行下一個階段的詞彙叢聚分析。

（二）詞彙叢聚分析

533個特徵詞中，不少詞彙由於經常出現在同一篇文章中，或是經常一起被使用，表示具相同意義與用途，不應視為兩種個體，因此本研究利用調整型phi係數（adjusted phi coefficient）計算詞彙相似度，並具以進行詞彙叢聚。

叢集辭彙數量多寡將影響文件分群效果（Baker & McCallum, 1998），在決定詞彙叢聚數量時，我們採用自動化二分式叢集法，當群集數上升，導致群集相似距離變小時，就不再分裂群集。各詞彙間相關係數矩陣如下表3所示：

叢聚過程示意圖如圖2所示，當相似距離小於0.653後，會導致群集分裂數值大幅增加，使得叢聚效果下降，本研究經過應用軟體執行得到118個詞彙叢集作為文件特徵詞彙叢聚資料庫，如表4所示。

表 3 特徵詞彙相關係數矩陣(部份)

	ABILITY	ACCOUNT	ACCURACY	ACCURATE	ACHIEVE	ACQUIRE	ACQUIRED
ABILITY	1	0.494	0.528	0.555	0.523	0.496	0.546
ACCOUNT	0.494	1	0.514	0.511	0.502	0.478	0.48
ACCURACY	0.528	0.514	1	0.558	0.545	0.461	0.481
ACCURATE	0.555	0.511	0.558	1	0.475	0.508	0.484
ACHIEVE	0.523	0.502	0.545	0.475	1	0.536	0.509
ACQUIRE	0.496	0.478	0.461	0.508	0.536	1	0.558

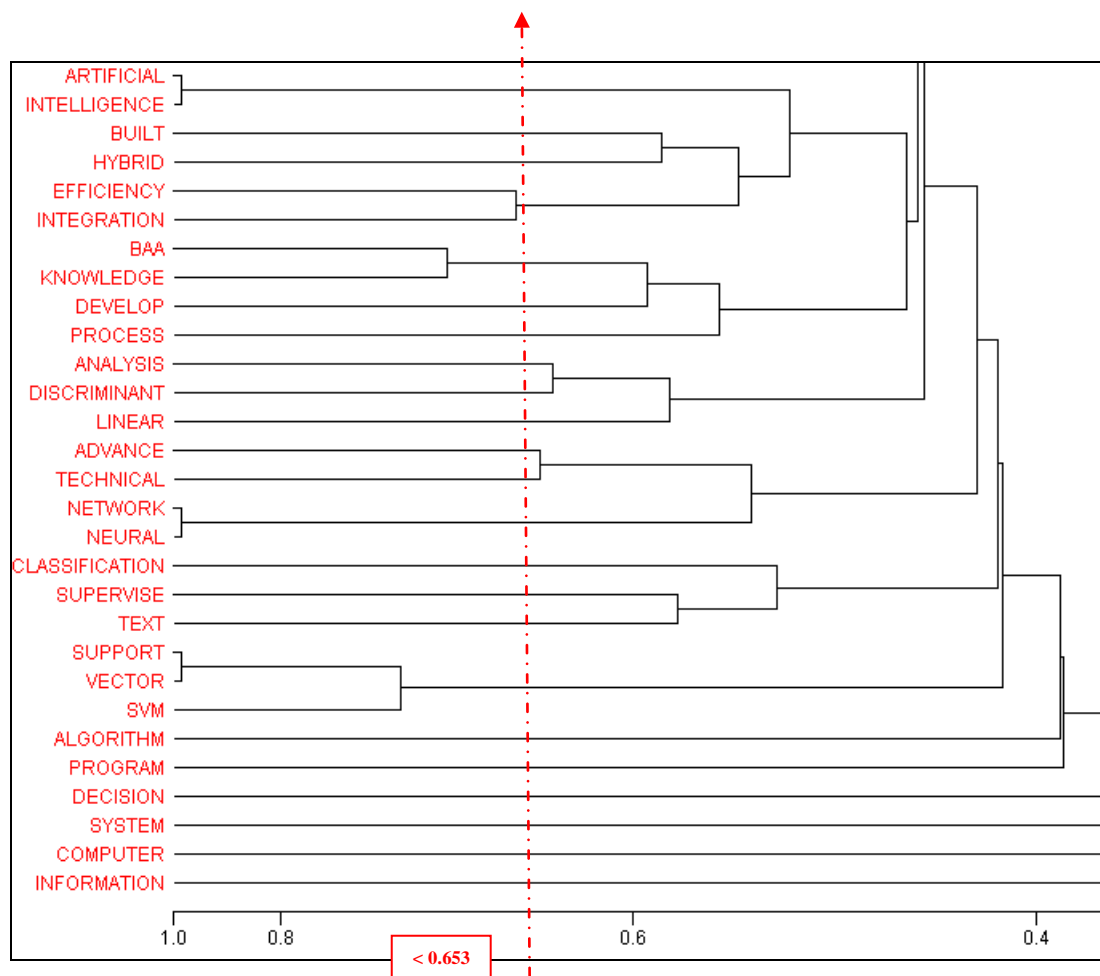


圖 2 詞彙叢聚過程示意圖

各詞彙叢集命名策略，則是依各叢集內含詞彙意義進行命名，在命名時主要以名詞為主，如 economic、financial、forecast、future、investigate、market 等詞彙集，就命名為 financial economic 群集，在命

名時也一併考慮本研究之研究主題為machine learning的未來趨勢分析為依歸，取其具研究方向相關意涵之名稱。

表4 詞彙叢集與命名

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NB CASES	% CASES	TF · IDF
NEURAL NETWORK	389	1.2%	0.7%	0.4%	96	17.3%	296.1
WEB/CONTEXT/DOCUMENT QUERY	687	2.1%	1.3%	0.7%	207	37.4%	293.7
LANGUAGE & LINGUISTICS	420	1.3%	0.8%	0.4%	111	20.0%	293.2
MODEL	645	2.0%	1.2%	0.7%	198	35.7%	288.2
OPTIMAL PROBLEM	1057	3.3%	1.9%	1.1%	296	53.4%	287.7
EDUCATION	386	1.2%	0.7%	0.4%	102	18.4%	283.7
ACCURACY PREDICTION	688	2.1%	1.3%	0.7%	222	40.1%	273.2
INFORMATION RETRIEVAL	651	2.0%	1.2%	0.7%	211	38.1%	272.9
WORD EXTRACT	582	1.8%	1.1%	0.6%	194	35.0%	265.2
TEXT CLASSIFICATION	432	1.3%	0.8%	0.4%	144	26.0%	252.8
HEALTHCARE/TREATMENT IDENTIFICATION	1094	3.4%	2.0%	1.1%	328	59.2%	249
SVM	403	1.2%	0.7%	0.4%	134	24.2%	248.4
KNOWLEDGE PROCESS	567	1.8%	1.0%	0.6%	218	39.4%	229.7
RULE INDUCE	369	1.1%	0.7%	0.4%	138	24.9%	222.7
.....							
HAND	28	0.1%	0.1%	0.0%	16	2.9%	43.1
STRUCTURED	27	0.1%	0.0%	0.0%	17	3.1%	40.9
CONDUCT	25	0.1%	0.0%	0.0%	23	4.2%	34.5
FINALLY	25	0.1%	0.0%	0.0%	25	4.5%	33.6
WIDE	23	0.1%	0.0%	0.0%	20	3.6%	33.2
HANDLE	21	0.1%	0.0%	0.0%	16	2.9%	32.3
DEAL	21	0.1%	0.0%	0.0%	20	3.6%	30.3

伍、文獻群集與機器學習領域的趨勢分析

一、群集分析

擷取出machine learning相關文獻的特徵詞彙後，本研究透過自組織映射網路，在無法得知最合適群集數的前提下，利用自組織映射網路（self-organizing map network, SOM）進行資料歸納，SOM是一種非監督式（non supervised）學習網路模式，在1980年由Kohonen提出，其主要目的將資料適當分群，作為辨識與分析不同群集其特徵屬性之基礎。在自組織映射網路中，拓撲層（topology level）的類神經元是以矩陣的方式排列，並且根據目

前的輸入向量，彼此競爭以爭取到調整網路加權值的機會，而最後輸出層的神經元會根據輸入向量的「特徵」以有意義的「拓樸結構」(topological structure) 展現在輸出空間中，由於所產生的拓樸結構圖可以反應輸入向量本身的特徵，因此稱作為自組織映射網路。

自組織映射網路的基本原理為利用大腦結構的特性，大腦中具有相似功能的細胞聚集在一起，如人類大腦中有專司視覺、聽覺、味覺等區塊的組織，也就是腦神經有「物以類聚」的特性，自組織映射網路也就是模仿這樣的特性，其輸出處理單元會相互影響，當網路學習完畢後，其輸出處理單元相鄰近者會具有相似的功能，也就是擁有相似度極高的連結權數。由於SOM的特性能自動將大量資料分成若干群集，並將相似度較高的資料聚集在一起，標記出每一群集的主要特徵屬性，最後分出最佳的群集數。本研究利用自組織映射網路能自我群集尋找最適合群集數的訓練過程，得到最佳的群集數，如表5所示。

表 5 網路參數配置

Input Layer	1,075 neurons
Output Layer	10 neurons
Input Field	118 fields
Instances	554 instances
Learning rate decay	Linear
Neighborhood	2
Initial Eta	0.1
Cycles	150

二、群集命名

群集命名則依各群集內所含特徵詞彙叢聚的內容進行文件群集命名，命名結果與各群集文件數如下表6所示。

表 6 分群結果與群集命名

群集編號	文件數	主要詞彙叢集(列舉)	群集命名
1	111	HEALTHCARE/TREATMENT INFORMATION RETRIEVAL WEB/CONTEXT/DOCUMENT QUERY	非結構化文件及網頁文字 辨識領域
2	38	RULE INDUCE ARTIFICIAL INTELLIGENCE ALGOTITHM	專家系統與人工智慧演算 法領域
3	23	LANGUAGE & LINGUISTICS FACE CAPTURE	自然語言與語意學及表情 辨識領域
4	77	FINANCIAL ECONOMIC NEURAL NETWORK ACCURACY PREDICTION	財務經濟預測與最佳化應 用領域

5	35	HEALTHCARE/TREATMENT HUMAN INTERACTION SYSTEM	醫療照護與人機互動系統 領域
6	29	DECISION/INDUSTRIAL MANUFACTURE SEQUENCE JOB/SINGLE	生產製造與工作排程領域
7	85	REPRESENTATION/PHENOMENON BEHAVIOR PSYCHOLOGY/KNOWLEDGE PROCESS	人類行為工程學與知識處理 領域
8	34	OPTIMAL PROBLEM KNOWLEDGE PROCESS SEQUENCE JOB/SINGLE MODEL	訊息處理與知識學習領域
9	30	EDUCATION HUMAN INTERACTION BEHAVIOR PSYCHOLOGY	教育及行為心理學領域
10	92	ERGONOMICS/CYBERNETICS KERNEL FUNCTION BEHAVIOR PSYCHOLOGY	人類大腦工程學與神經信號 處理領域

陸、機器學習領域各學門分析

各領域文獻自1991年後，每年成長趨勢繪製時間趨勢折線圖，分別如圖3~圖12所示，可知各領域文獻之成長趨勢，近年來各研究領域的應用與文獻發表都有逐年成長現象。

非結構化文件處理及網頁文字辨識領域於2000年後有呈持續上升的現象；自然語言與語意學及表情辨識領域除了於2000~2004有下降有情況，自2005後亦為持續成長；專家系統與人工智慧演算法應用領域則波動較大，自2005年後波動而且有漸漸減緩的現象；財務經濟預測與最佳化應用領域除2004年間有下降情況，其餘均緩慢上升；醫療照護與人機互動系統領域自2006年後呈大幅度上升的情況；生產製造與工作排程領域與人類行為工程學與知識處理領域大致來看都呈穩定成長的現象；訊息處理與知識學習領域自2005年後，有顯著上升的情況；教育及行為心理學領域則自1999年後有成長趨緩的情況，但2008年有較多的研究發表；人類大腦工程學與神經信號處理領域除2007外，亦為緩慢成長的現象。

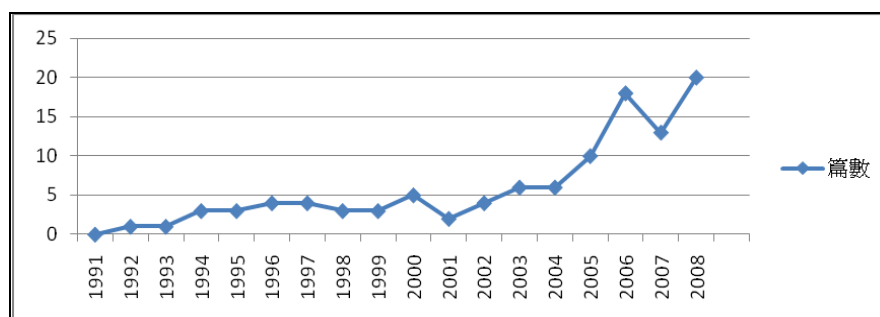


圖3 非結構化文件處理及網頁文字辨識領域趨勢分析圖

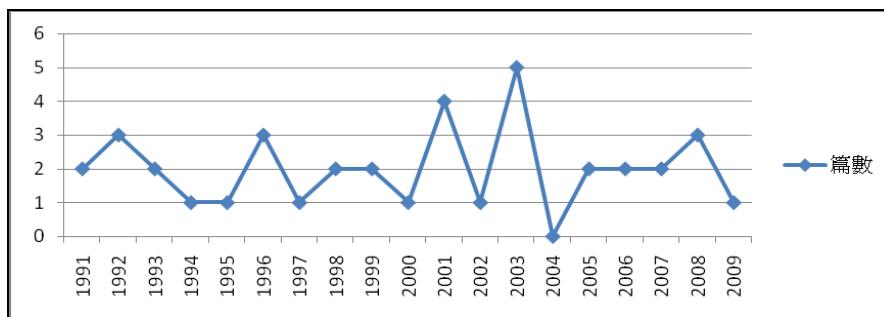


圖 4 專家系統與人工智慧演算法應用領域

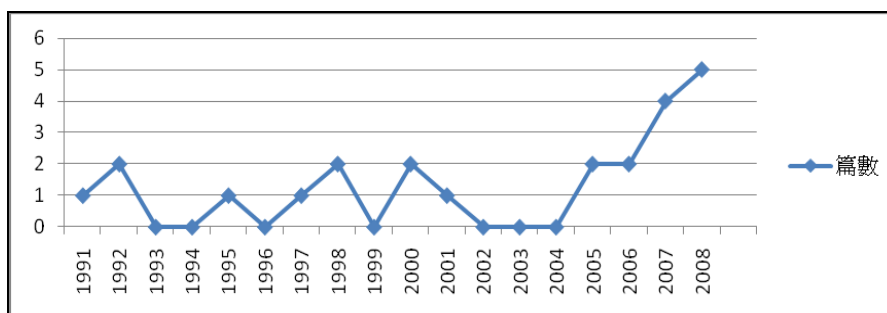


圖 5 自然語言與語意學及表情辨識領域

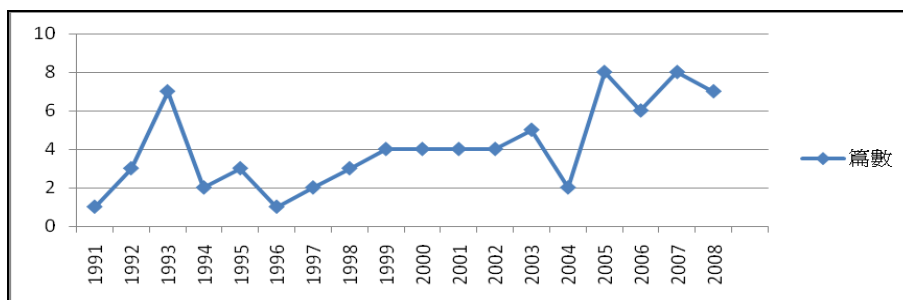


圖 6 財務經濟預測與最佳化應用領域

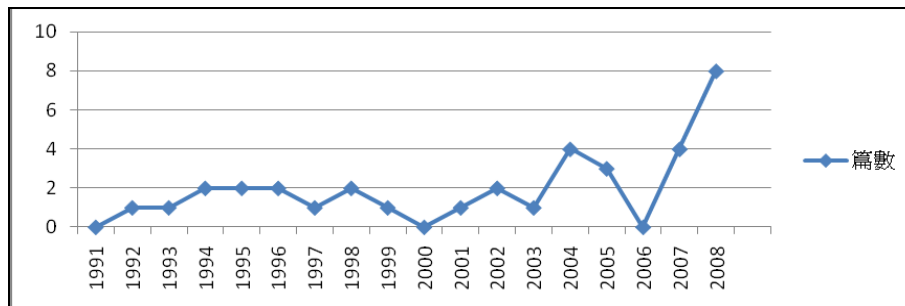


圖 7 醫療照護與人機互動系統領域

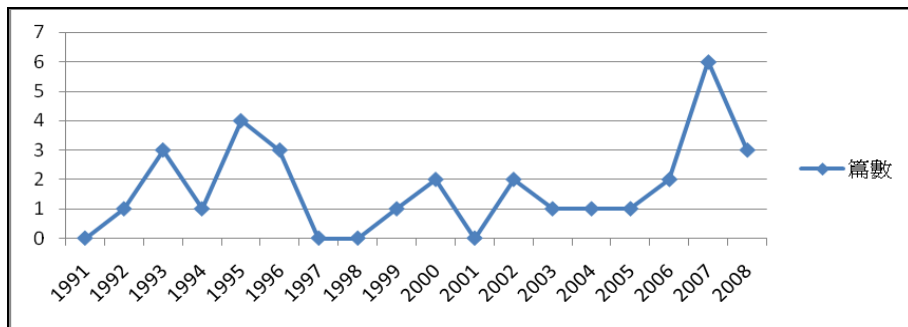


圖 8 生產製造與工作排程領域

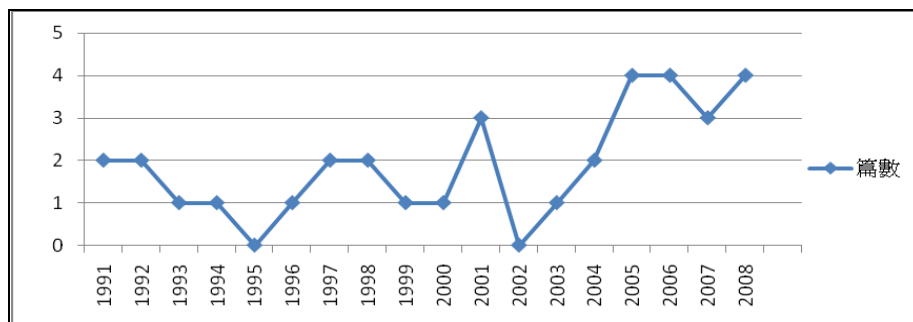


圖 9 人類行為工程學與知識處理領域

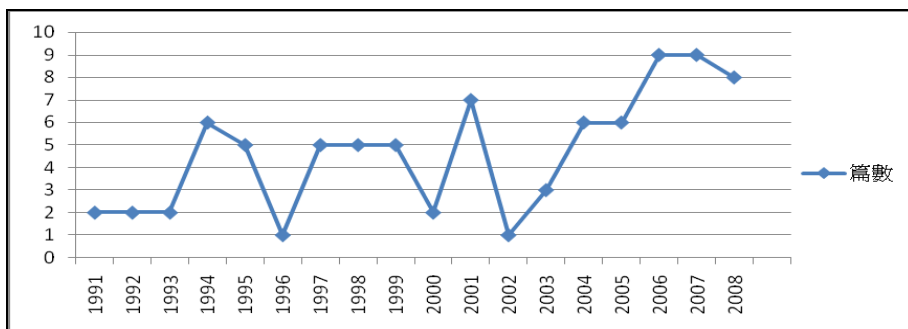


圖 10 訊息處理與知識學習領域

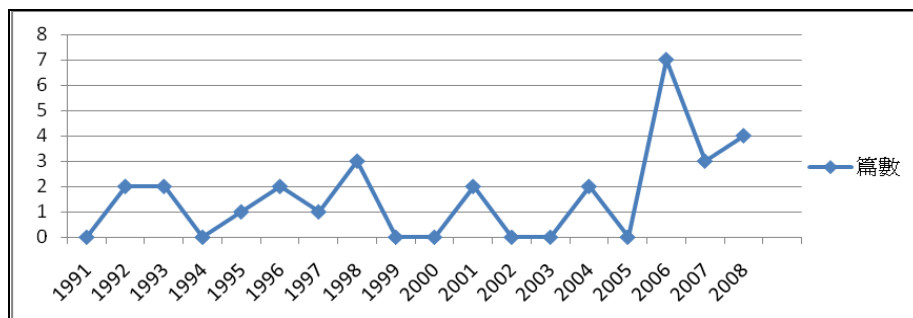


圖 11 教育及行為心理學領域

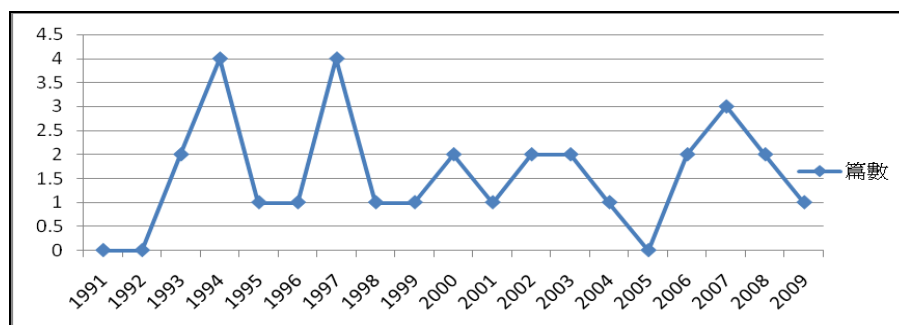


圖 12 人類大腦工程學與神經信號處理領域

柒、結論與未來研究方向

象徵人類研究的期刊創作與發表，一向是未來科技趨勢的最佳寫照，本研究成功的利用期刊文獻的關鍵欄位的文字進行文字擷取、詞彙叢聚與文件分群，並結合文獻數、發表年份進行機器學習應用領域的趨勢分析，透過趨勢圖，不僅能正確的解讀各研究領域的歷史及發展趨勢，以利研究人員觀察機器學習方法應用領域的發展脈絡，同時也提出一個基於文字勘探的趨勢分析模型，未來亦可應用於其它領域科技的趨勢研究。

而本研究主要針對SSCI資料庫，因此對於醫療、資訊工程、理工等研究領域的SCI文獻資料庫無法取得，未來若能擴大文獻資料樣本，一併考量SCI資料庫機器學習的期刊文獻，納入更多詞彙庫數量，做更進一步的分析與解讀。未來更可配合研究重大研究新發現的年份、世界重大事件發生的年份（國家策略、金融危機）等構面進行樞紐分析，以找出影響每一個研究領域成長趨勢的可能因素。

參考文獻

- Aas, K., & Eikvil, L. (1999, June). *Text categorization: a survey* (Report No. 941). Norwegian Computing Center. U.S.A.
- Baker, L. D., & McCallum, A. K. (1998). *Distributional clustering of words for text classification*. Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 96-103.
- Bassiou, N., Kotropoulos, C., & Pitas, J. (2001). *Hierarchical word clustering for relevance judgment in information retrieval*. In Conjunction with the Third International Conference on Enterprise Information Systems, 139-148.
- Bekkerman, R., El-Yaniv, R., Winter, Y., & Tishby, N. (2001). *On feature distributional clustering for text categorization*. Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval, 146-153.
- Davis, L. D., & Mitchell, M. (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York.
- Garcia, A. J. J., Pikatza, J. M., Florez, S., & Sobrado, F. J. (2005). *Intrusion detection using text mining in a web-based telemedicine system*. Proceedings of the 18th

- Australian Joint Conference on Artificial Intelligence.
- Goldberg, D. E. (1989). *Genetic Algorithm in Search, Optimization and Machine Learning*. Addison-Wesley, New York.
- Heller, K., & Ghahramani, Z. (2005). *Bayesian hierarchical clustering*. ACM International Conference Proceeding Series; Vol. 119, Proceedings of the 22nd international conference on Machine Learning, 297-304.
- Kao, A., & Poteet, S. (2006). *Text Mining and Natural Language Processing – Introduction for the Special Issue*. Springer-Verlag, New York.
- Lam, W., Ruiz, M., & Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 865-879.
- Lu, W., Chien, L., & Lee, H. (2002). *Translation of Web Queries Using Anchor Text Mining*. ACM Transaction on Asian Language Information Processing (TALIP), 1(2), Issue 2, 159-172.
- Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, 33, 1455-1465.
- Moens, M. F., & Dumortier, J. (2000). Text Categorization: the Assignment of Subject Descriptors to Magazine Articles. *Information Processing & Management*, 36, 841-861.
- Moretti, S. (2006). *Minimum Cost Spanning Trees Situations and Gene Expression Data Analysis*. ACM International Conference Proceeding Series, Vol. 199, Proceedings from the 2006 workshop on Game theory for Communications and Networks.
- Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Process and Management*, 24(5), 513-523.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Survey*, 34(1), 1-47.
- Sebastiani, F. (2005). *Text Categorization, Text Mining and its Applications*, WIT Press, Southampton, U.K., 109-129.
- Stumme, G., Hotho, A., & Beremdt, B. (2002). *Usage mining for and on the semantic web*. The Semantic Web – ISWC 2002, 1st International Semantic Web Conference, Lecture Notes in Computer Science, 2342, 264-278.
- Yang, Y., & Pedersen, J. (1997). *A comparative study on feature selection in text categorization*. Proceedings of the 14th International Conference on Machine Learning, ICML-97, 412-420.