Support Vector Machines 自動分類技術 應用於網路文件間之相關性量測

李俊宏[1] 楊正宏[1] 徐豐智[2] 陳廷忠[2]
國立高雄應用科技大學 電子計算機中心[1]
國立高雄應用科技大學 電機工程研究所[2]
leechung@mail.ee.kuas.edu.tw[1]
chyang@cc.kuas.edu.tw[1]
{fonze0101, tinchung } @dml.ee.kuas.edu.tw[2]

摘要

本研究提出一個以自動文件分類技術為基礎的多重分類器架構，結合在文件分類領域中有顯著效能的 Support Vector Machines(SVMs)演算法進行網路文件間之語意相關性量測 (Semantic Relatedness Measure)之運算平台建立，將網路文件原始之詞彙特徵向量透過已經訓練好的 SVM 分類器做決策後產生新的語意向量。實驗結果顯示透過量測文件語意向量之間的距離(Distance)餘弦函數(Cosθ)、Dice 及 Jaccard，可以得到文件間語意相關性量化值。

關鍵詞：Support Vector Machines、文件自動分類、文件探勘、機器學習。

Abstract

In this paper we present a novel measuring method using a multi-classifier platform to perform evaluation of semantic relatedness among texts. We employed several text classifiers based on various specific topics using support vector machines (SVMs) to construct a multi-classifier platform. Firstly, we employ our developed algorithm to deal with text pre-processing and training for classifier generation. Subsequently, the texts of unknown category go through the trained SVM classifiers to generate new vectors of decision features made by the classification results. Essentially, the resulting class vectors are used to represent semantic vectors of respective texts for comparison of relatedness with other texts. In addition, we evaluated the system performance with some traditional textual similarity evaluation techniques, including Distance, Inner, Cosine and Dice methods.

Keywords: Support Vector Machines、Text Categorization、Text Mining、Machine Learning

1. 前言

隨著網際網路的興起及文件普遍電子化的趨勢下，網際網路上的資料量每秒鐘以倍數的方式成長，整個網際網路上充斥著成千上萬半結構化及未結構化的文件，對於使用者而言，如何在面對如此龐大而繁雜的資料集合中找到符合使用者需要的資料是個極具挑戰性的問題，近年來也有許多國內外學者投入進行相關的研究。

近年來由於資料探勘(Data Mining)及文件探勘(Text Mining)技術的蓬勃發展，透過不同的探勘技術如：分類(Classification)、分群(Clustering)等，可以將龐大文件集合中文件間隱含的資訊有效的挖掘出來，然而經由這些技術的挖掘固然可以呈現文件間隱含的資訊，卻無法提供給使用者知道文件間內容的相關程度，若藉由人工的方式來進行判定文件間內容的相關程度，會因每個人主觀的判斷而有所誤差，最理想的方式是能藉由一個量化的數值來讓使用者瞭解文件間相關程度的強弱關係。此技術可以應用於搜尋引擎、知識挖掘、文件探勘等相關領域上，讓使用者面對網際網路上龐大的資料集合時可以找到真正符合需要的資訊。

傳統的文件分類多半藉由判定文件間『相似度』作為分類的依據，經由計算後將相似程度高的文件集合於某一個『文件叢集』中，此種方法廣泛用於文件檢索系統的理論模型設計；『相似度量測(similarity measurement)』亦是資訊檢索過程中不可或缺的一部分。『語意相關性(semantic relatedness)』相較於『語意相似性』是更為廣泛的概念；語意相似度只是代表語意相關度的一種應用，Resnik[9]試著以一個例子說明相似度與相關度的差異性：以相似度的觀點來看，“汽車與汽油”之間的關係似乎比“汽車與腳踏車”來的低；但若以相關度的觀點來說，前者又較後者來高；對大多數計算應用的系統而言，相關度應用的場合遠高於相似度。

隨著搜尋引擎近年來的發展越來越多元化，越來越多的搜尋引擎推陳出新，若可透過一個語意相關性模型量測出兩網路文件間之語意相關性並透過數值的方式呈現，此技術可以應用於搜尋引擎技術上的提升，藉由文件間量化數值作為搜尋相關網路文件的依據之一，以有效的減少搜尋引擎搜尋結果中冗餘的網路文件。此技術未來發展亦可進一步應用於網路上相關多媒體文件的語意相關性量測。

計畫編號:NSC93-2213-E-151-005

2.相關文獻探討

本研究的目標是透過 SVMs 多重類別分類技術為基礎建立一個運算平台，經由每個分類器的決策值產生新的語意向量，透過文件間語意向量可計算出一個量化的數值來表示文件間語意相關程度的強弱。下面的章節中將介紹數個量測語意相關程度及相似程度的方法。

2.1 Latent Semantic Indexing(LSI)

Latent Semantic Indexing (LSI)是由 Deerwester 於 1990[4] 提出的一個資訊檢索(Information Retrieval, IR)的模型，其與傳統向量空間模型(Vector Space Model)的差別如圖 2.1 所示。傳統向量空間模型藉由關鍵字集合與文件集合所組成的二維矩陣來呈現，LSI 透過奇異值分解(Singular Value Decomposition, SVD)將傳統向量空間模型轉換成 LSI 向量空間模型(或稱語意空間)。在 LSI 向量空間模型上文件及詞彙將在同一特徵空間上呈現，可透過語意相似度量測演算法量測文件間，詞彙間，詞彙與文件間之語意相似程度。當使用者透過詞彙查詢(query)，LSI 在語意空間上透過運算餘弦函數(cosine) 將查詢詞彙與文件間運算後相似性最高的文件檢索出來。

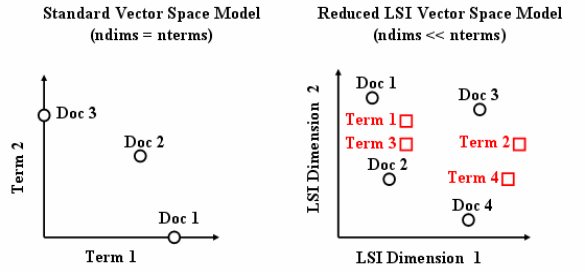

圖 2.1 傳統向量空間模型與 LSI 向量空間模型之比較[5]

2.2 Latent Semantic Analysis(LSA)

Latent Semantic Analysis (LSA) 是由 Landauer 於 1997[7]所提出之語意相似度分析模型，以圖 2.2 為例。由文件及詞彙所組成行 12 列 9 的原始矩陣{X}透過奇異值分解(Singular Value Decomposition, SVD)成三個矩陣，{W}為行 12 列 9 以列向量(Row Vector)對應詞彙在語意空間(semantic space)所呈現方式，{S}為 9 個元素之對角矩陣代表語意空間，$\{P^T\}$為行 9 列 9 以列向量對應文件在語意空間所呈現方式。選定縮減維度的值為 2，將經由奇異值分解之矩陣轉換並相乘產生新的語意空間模型。原始矩陣由文件與詞彙所組成僅能看出詞彙在每篇文件出現的頻率，經由 LSA 模型透過 SVD 產生新的語意空間模型，並可透過餘弦函數(cosine)來計算詞彙間語意相似程度。

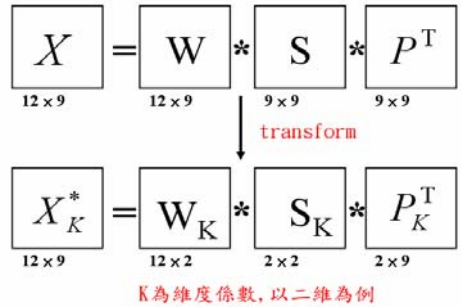

圖 2.2 LSA 奇異值分解示意圖

2.3 WordNet

WordNet 是一個詞彙參考資料庫，構成 WordNet 主體並非是詞彙本身，而是詞彙所蘊含的意義，所有詞彙組成同義字集合(Synonym Sets, Synsets)，每個集合代表一個基本的詞彙概念(Concept)，其中由美國普林斯頓大學所發展的「WordNet」系統[6]以初步具備概念的結構，而且有多種意義的詞彙會同時出現在不同的同義字集合中，同義字集合之間以不同的關係連結。以英文名詞為例，在 WordNet 中定義了四種關係[11]：

1. Synonym & Antonym 同義詞與反義詞
2. Hypernym & Hyponym 上位詞與下位詞
3. Holonym (relation is part of) 完全關係
4. Meronym (relation part of) 附屬關係

上述三種方法，量測文件或詞彙之間語意相關程度；以 WordNet 架構而言，比較文件或詞彙之間所在之概念集合重疊度(Overlap)或計算之間的階層數來表現兩文件或詞彙之間相關程度的強弱，但卻無法用數值精準的表現文件或文字間相關程度。LSI 與 LSA 透過奇異值分解將文件與詞彙間，及詞彙與詞彙間隱含的語意呈現出來，並量測相似程度。本研究希望能在不借助 WordNet 特殊架構及不同於 LSI 與 LSA 語意相似度量測，直接進行兩文件間語意相關程度量測，並用數值表現兩文件間語意相關程度。

3. Support Vector Machines 相關技術

Support Vector Machines (SVMs)是由 Vapnik 及其團隊於 AT&T 貝爾實驗室中所發展出來，其起源於統計學習法則中的結構風險最小化(Structural Risk Minimization：SRM)[10]，Vapnik[12]透過統計的方式證明 SVMs 在資料趨近無限大時，亦可以在有限的次數中找到最佳解。SVMs 最初的設計是處理二元分類的問題(Binary Classification)，經由運算兩類別樣本空間的最佳分割超平面(Optimal Separate Hyper plane)以確保最小錯誤分類率；處理線性不可分割的分類問題上，SVMs 將在原始樣本

空間無法分割的樣本映射至高維度的特徵空間中進行分割，或者是導入柔性邊界(Soft Margin)機制，允許若干個樣本在訓練階段可以錯誤分類，將原始樣本空間中無法線性分割二類訓練樣本的問題轉化成可以線性分割。處理不可線性分割問題也是 SVMs 另一個優點。

3.1 線性可分割問題

首先我們必須先假設存在一組訓練樣本集合 S 如下：

$$S = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_l, y_l)\} \quad (3.1)$$

$$x_i \in R^n, y_i \in \{+1, -1\}, \ for \ \ i = 1 \sim m$$

其中 m 代表訓練集合 S 之樣本個數，n 為訓練樣本向量之維度，存在一超平面(Hyperplane)能將二類訓練樣本完全分隔，該平面描述為：

$$f_H(x) = w \cdot x + b, \quad (3.2)$$

因此我們可根據(3.3)式之決策函數(Decision Function)將二類訓練樣本 x_i 分隔，

$$f_D(x_i) = sign(w \cdot x_i + b) = \begin{cases} +1, & if \ \ y_i = +1, \\ -1, & if \ \ y_i = -1, \end{cases} \quad (3.3)$$

此處 $w \in \Re^n$ 且 $b \in \Re$；如果存在 (w, b) 使得所有的樣本 x_i 均滿足(3.4)不等式之情況下，我們可以將此訓練樣本集合 S 稱作可被線性分割。

$$f_D(x_i) = w \cdot x_i + b \ \begin{cases} \geq +1, & if \ y_i = +1, \\ \leq -1, & if \ y_i = -1, \end{cases} \quad i = 1,2,3,\cdots,m \quad (3.4)$$

SVM 分類器沿著超平面的垂直方向將二類別邊界(Margin：$\rho(w,b)$)擴展至最大，使得分類錯誤可能性降到最低，得到唯一最佳解，如圖 3.1 所示。

$$\rho(w,b) = \min_{\{x|y=1\}} \frac{w \cdot x + b}{\|w\|} - \max_{\{x|y=-1\}} \frac{w \cdot x + b}{\|w\|} = \frac{2}{\|w\|} \quad (3.5)$$

其中 w 代表超平面法向量(Normal Vector)，b 代表超平面之偏移量(bias)，將(3.5)式中最大邊界問題表現如下式：

$$\text{minimize} \ \ \frac{1}{2} w \cdot w$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1, \ i = 1,2,3,\cdots,m \quad (3.6)$$

因此，尋找唯一最佳超平面是典型的二次規劃(Quadratic Programming：QP)問題，可由 Lagrangian 乘式法求解，將(3.6)式問題轉化成：

$$\text{maximize} \ \ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$\text{subject to} \ \begin{cases} \sum_{i=1}^{m} y_i \alpha_i = 0 \\ \alpha_i \geq 0, \ i = 1,2,3,\cdots,m \end{cases} \quad (3.7)$$

針對這類限制最佳化問題於求解與分析中，根據 Karush-Kuhn-Tucker (KKT)定理可以解決(2.7)式問題，成功的求得一組解 (w, b, α)。

$$\bar{\alpha}_i (y_i (\bar{w} \cdot x_i + \bar{b}) - 1) = 0 \quad , \quad i = 1, 2, 3, \cdots, m \quad (3.8)$$

為了滿足(2.4)不等式，在上式中 $\bar{\alpha}_i$ 必須是非零變數，且相對應的樣本向量 x_i 滿足(3.8)式既稱為支撐向量(Support Vector：SV)。為了構成最佳超平面 $(\bar{w} \cdot x + \bar{b})$，隨著(3.8)式可得下式

$$\bar{w} = \sum_{i=1}^{m} \bar{\alpha}_i y_i x_i \quad , \quad (3.9)$$

而且偏移量 $\bar{b}$ 可以透過 KKT 條件(3.8)式決定

$$\bar{b} = y_i - w \cdot x_i \quad (3.10)$$

最後可將原本的最佳化超平面重新定義成下式：

$$f_H(x) = \sum_{i=1}^{m} \bar{\alpha}_i y_i x_i \cdot x + \bar{b} \quad , \quad (3.11)$$

而決策函數則改寫成

$$f_D(x) = sign(f_H(x)) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i x_i \cdot x + \bar{b}) \quad . \quad (3.12)$$

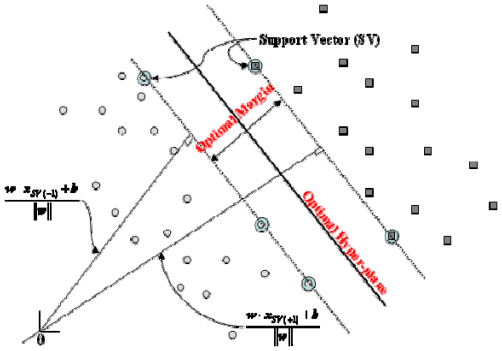

圖 3.1 SVMs 分類問題之架構

3.2 線性不可分割問題

在現實生活的應用中並非所有的問題都可以順利的透過線性分割來分類，對於無法進行線性分割的問題 SVMs 提供下列兩種方式進行分割：

(1) 柔性邊界(Soft Margin)
(2) 非線性核心函數(Non-Linear Kernel Function)

3.2.1 柔性邊界(Soft Margin)

如果樣本集合 S 在原始樣本空間上無法進行線性分割，在不更改原始空間的 SVMs 分類器之原則下導入一鬆弛變數(Slack Variable：$\xi_i \geq 0$)於(3.4)式中，其修改後不等式如下：

$$f_D(x_i) = w \cdot x_i + b \ \begin{cases} \geq +1 - \xi_i, & if \ y_i = +1, \\ \leq -1 - \xi_i, & if \ y_i = -1, \end{cases} \quad i = 1,2,3,\cdots,m \quad (3.13)$$

ξ_i 表示編號 i 的訓練樣本之誤差(如圖 3.2)，SVMs 利用一調和係數來控制被允許 ξ_i 的大小，經過修正

後的數學模型如下：

$$\text{maximize} \quad \sum_{i=1}^{m}\alpha_i - 2^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_j^{T} x_i \ ,$$

$$\text{subject to} \ \sum_{i=1}^{m}\alpha_i y = 0 \quad \text{and} \quad 0 \le \alpha_i \le C \quad \text{for all} \qquad (3.14)$$

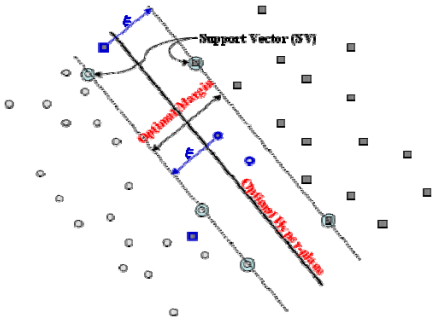

圖 3.2 鬆弛變數 ξ 導入 SVMs 之情況

3.2.2 非線性核心函數(Non-Linear Kernel Function)

除了導入鬆弛變數(Slack Variable)將邊界變的更有彈性外，SVM 分類器亦可以透過向量映射至特徵空間(Feature Space)，將原本無法在原始空間(Original Space)進行線性分割情況之問題，利用映射到更高維度之特徵空間中進行線性可分割問題，SVMs 提供 Kernel Function $K(\cdot,\cdot)$ 可將繁瑣複雜的特徵映射程序簡化，不用運算特徵空間中所映射向量之個別維度，只須利用原始空間的向量代入 Kernel Function 運算出對映在特徵空間上向量之內積運算。接著將訓練後所選擇的支撐向量(Support Vectors)來決定最佳二類分割超平面並對未分類之資料進行決策與分類。

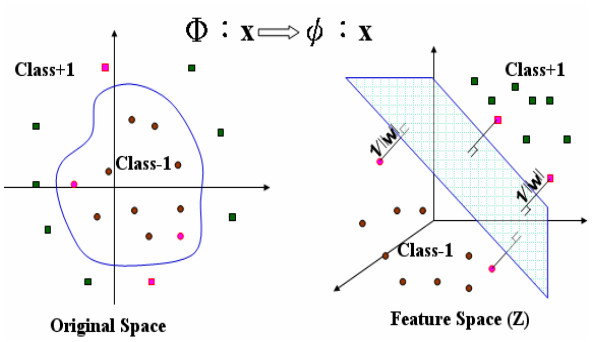

圖 3.3 原始空間映射至特徵空間之示意圖

4. 實驗架構

經由相關文獻的探討與研究後，擬定主要的研究方向為透過多類別分類的方法將文件間語意相關程度透過量化的數值來表現其強弱，如圖 4.1 所示，在實驗架構上分為兩大階段來進行，第一階段為分類器的設計及訓練，第二階段為語意相關性量測模型的設計。

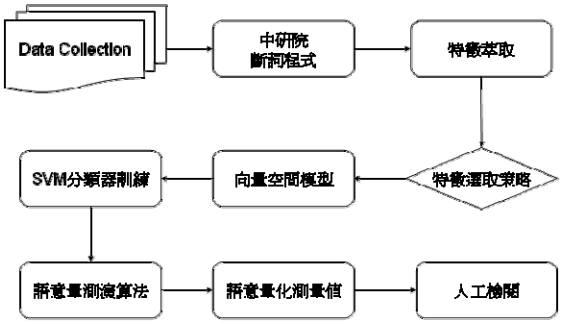

圖 4.1 本研究實驗流程圖

4.1 語料庫

中文文件分類研究上由於沒有像英文文件分類上有標準的資料集合提供研究使用，本研究語料庫利用自行收集的中文文件資料建立中文文件語料庫，基於資料多樣性的考量，在蒐集文件資料時透過不同的新聞網站來進行蒐集（如表 1），為避免所蒐集的新聞特定集中在某些特定的時間點上，故本研究之語料庫仍持續增加中以增加實驗的強韌性。

表 1 文件資料來源網站表

資料來源 網站名稱	網址
YAHOO	http://www.yahoo.com.tw
PC-home	http://www.pchome.com.tw
中央日報	http://www.cdn.com.tw
台灣時報	http://www.taiwandaily.com.tw
聯合新聞網	http://www.udn.com
中時電子報	http://news.chinatimes.com

4.2 資料模型處理程序

首先將蒐集的新聞文件依照類別加以分類，透過中研院開發出來的斷詞程式進行斷詞，接著利用本研究自行開發的詞彙擷取程式配合實驗的特徵選取策略將所需的文件特徵萃取出來，將萃取出來之文件特徵組成特徵向量並作為文件代表的模型，經由運算得到每一個特徵之權重值並紀錄於文件模型中作為訓練集合及測試集合之文件向量模型。

4.3 SVMs 分類器之訓練

在 SVMs 訓練階段，將訓練資料與驗證資料送入 SVMs 分類器中，根據本實驗室先前的相關研究[1,2,3]決定選用 Gaussian RBF 核心函數(Kernel

Function)，並配合設定參數(包含調和係數 C、核心函數之參數)與結束條件。利用 SVMs 最佳化演算法，例如 Projection、連續最小最佳化(Sequential Minimal Optimization：SMO)等，藉由調整所有訓練文件之權重值求得最佳決策函數，即最佳分割超平面，在訓練完成後將所有訓練文件之權重值輸出作為測試階段決策函數建立之參數。

4.4 語意向量轉換

本研究利用數個類別來模擬現實生活中的文件類別作為語意向量之特徵，每篇文件均各別經由不同類別 SVMs 分類器決策出一數值(如圖 4.2)形成一類別語意向量。

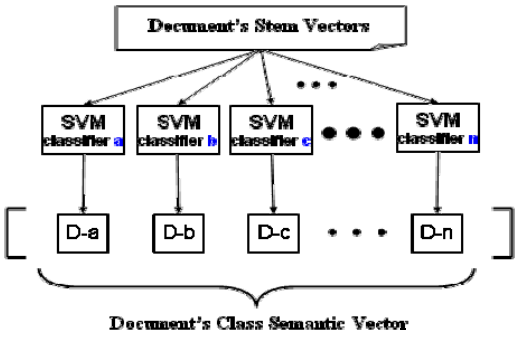

圖 4.2 文件語意向量架構

根據決策值的格式可分為兩種，第一種方式表示每一個分類器在決策後經過正規化將輸出只由「+1」與「-1」表現，文件語意向量只由+1 與-1 組成，此種方式只能呈現出兩文件是否同時屬於數個類別；第二種方式則直接將分類器之決策值透過 Symmetric Saturating Linear 函數正規化後，作為文件語意向量特徵之加權值，如圖 4.3 為經過 Symmetric Saturating Linear 函數正規化後可能之決策值，文件決策值若為+1 時(位置 a)，表示該文件完全屬於 SVMs 分類器之+1 類別；若決策值為介於 0 與+1 之間時(位置 b)，表示該文件某種程度屬於 SVMs 分類器之+1 類別；文件決策值若為-1 時(位置 c)，表示該文件完全屬於 SVMs 分類器之-1 類別，或是完全不屬於+1 類別；若決策值為介於 0 與-1 之間時(位置 d)，表示該文件某種程度屬於 SVMs 分類器之-1 類別，或是某種程度不屬於+1 類別。

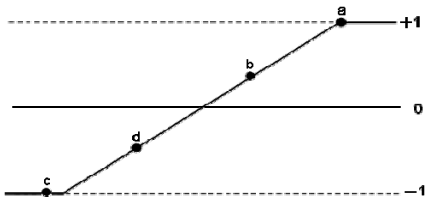

圖 4.3 SVMs 分類決策可能值

4.5 語意相關性量測設計

本研究透過多類別分類的架構將原始文件詞彙特徵轉變成文件間類別語意向量，藉由相關演算法量測兩篇文件的類別語意向量，並透過量化的方式呈現其語意相關程度。除了透過 SVMs 分類器作為本系統向量轉換的工作外，假設所有文件均可同時屬於不同類別，而每一個 SVMs 分類器所決策出來的結果均代表文件在該類別主題概念中所之表現程度。

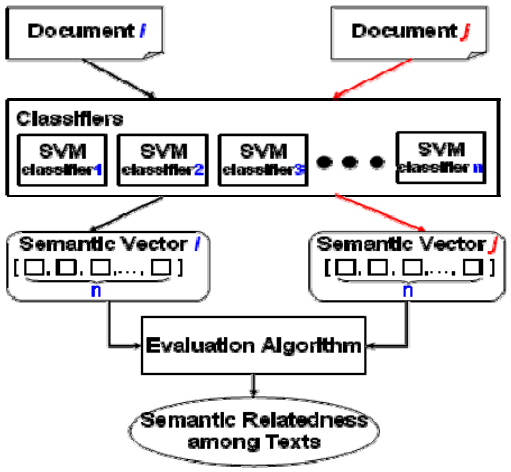

圖 4.4 兩文件語意量測架構

根據數個已訓練完成的分類器對測試文件 *i* 與文件 *j* 所決策判斷的結果做為此文件配對進行語意相關量測時之文件語意向量代表，進一步利用相關演算法量測兩篇文件語意向量間相關程度，最後將運算所得之數值正規化，即為兩文件間語意相關程度之量化值。

5. 結果與討論

若單純透過系統四種決策演算法決策出來的數值來判斷文件間之語意相關程度似乎又顯得不夠客觀，本實驗參考 Resnik 提出之方法，透過人工評量方式來做一個驗證。將文件間之相關程度分為五個等級，0%~20%、21%~40%、41%~60%、61%~80%與 81%~100%分別用數字 1~5 表示，將計算數個人的評量結果取平均值做為人工評量之依據。人工評量的結果也將與系統所計算四種量測方式運算之結果進行比較。下列實驗結果中，文件的編號以英文字母代表文件類別，數字代表該文件於該類別的編號。表 2 呈現測試文件中以政治為主題的文件與其它主題的文件透過四種不同的量測方式及人工評量的方式做一個比較。表 3 呈現測試文件中以影視為主題的文件與其它主題的文件透過四種不同的量測方式及人工評量的方式做一個比較。

表 2　測式文件之間的相關性量測值(一)

Measure of Semantic Relatedness between Documents					
Document pairs (以 document ID表示)	Distance	Cosine	Dice	Jaccard	Human Evaluation
b01 vs. a02	73.1%	90.7%	79.6%	71.1%	3
b01 vs. a04	77.8%	85.1%	79.5%	71.0%	4
b01 vs. b02	86.4%	95.9%	90.9%	84.7%	5
b01 vs. b04	96.2%	99.0%	98.8%	97.6%	5
b01 vs. c02	76.9%	69.7%	68.6%	61.4%	3
b01 vs. c04	71.9%	79.5%	73.1%	65.0%	3
b01 vs. d02	85.6%	80.4%	80.3%	71.8%	4
b01 vs. d04	75.7%	62.4%	61.9%	56.7%	3
b01 vs. e02	63.1%	40.4%	41.8%	46.2%	2
b01 vs. e04	70.6%	71.0%	67.2%	60.4%	3

表 3　測式文件之間的相關性量測值(二)

Measure of Semantic Relatedness between Documents					
Document pairs (以 document ID表示)	Distance	Cosine	Dice	Jaccard	Human Evaluation
e01 vs. a02	54.2%	55.1%	53.6%	51.9%	3
e01 vs. a04	58.3%	46.5%	48.5%	49.3%	2
e01 vs. b02	59.6%	47.4%	45.3%	47.8%	2
e01 vs. b04	69.7%	59.4%	57.9%	54.3%	3
e01 vs. c02	71.4%	68.3%	69.1%	61.8%	3
e01 vs. c04	61.7%	61.0%	62.5%	57.1%	3
e01 vs. d02	64.6%	37.0%	37.8%	44.6%	2
e01 vs. d04	57.5%	31.3%	28.1%	41.0%	2
e01 vs. e02	85.5%	92.8%	93.6%	88.7%	5
e01 vs. e04	89.9%	98.7%	97.2%	94.7%	5

經由實驗中發現，經由五個類別主題的 SVM 分類器轉換後的文件語意向量，其兩文件配對所運算出四種量測值與人工評估出的文件配對相關性量測值相似，最大誤差約為 1 個等級左右。在實驗中亦發現文件詞彙特徵向量轉換成文件類別語意向量此過程在本系統的文件相關性分析中時間花費最多。

6. 結論

本研究最大貢獻在於提出新穎的語意向量空間(Semantics-based Vector Space)模型，有別於透過辭彙特徵向量量測文件間相似性的研究方法，經由 SVMs 多重分類器系統之決策，將文件詞彙特徵向量轉換成文件類別語意向量，透過計算兩向量之距離或餘弦等，表現兩文件間相關(Relatedness)程度。經由實驗中証實經過 SVMs 轉換而得之文件類別語意向量所測量出文件相關程度之四種量化值(包括 Distance、Cosine、Dice 與 Jaccard)，均可明顯區分文件之間的主題是否相同，初步證實本研究所提之文件類別語意向量架構可行性。從數個語意量測的實驗中，發現我們所提出以 SVMs 分類器類別為特徵的類別語意向量可應用傳統向量測量的方法進行相關性的量化表現，其中又以透過餘弦量測方法效果最佳，透過距離量測方法效果最差。

參考文獻

[1] 李俊宏、李伯毅、徐豐智,2004。Support Vector Machines 應用於網路文件自動分類,2004 台灣網際網路研討會,pp.298-301,台東。

[2] 李俊宏、李伯毅、徐豐智, 一個以 Support Vector Machines 為主之中文文件自動分類系統的建構與特徵選取策略之分析，Journal of National Kaohsiung University of Applied Sciences,vol.2, pp.67~89, 2005.

[3] 李柏毅，Support Vector Machines 技術應用於中文文件自動分類之探討,國立高雄應用科技大學碩士論文,2004。

[4] Deerwester, S., Dumais, S., Furnas, G., Landauer, T.K., and Harshman, R., "Indexing by Latent Semantic Analysis." Journal of the American Society of Information Science, Vol.41 (6):pp.391-407, 1990.

[5] Dumais, S. T., Landauer, T. K. and Littman, M. L., "Automatic cross-linguistic information retrieval using Latent Semantic Indexing." In SIGIR'96-Workshop on Cross-Linguistic Information Retrieval, pp. 16-23, August 1996.

[6] Fellbaum, C., "WordNet: An Electronic Lexical Database", MIT Press. 1998.

[7] Landauer, T. K., & Dumais, S. T., "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge." Psychological Review, Vol.104, pp.211-240, 1997.

[8] Landauer, T. K., Foltz, P. W., and Laham, D., "Introduction to Latent Semantic Analysis." Discourse Processes, Vol.25, pp. 259-284, 1998.

[9] Resnik, P., "Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language". Journal of Artificial Intelligence Research, Vol.11, pp.95-130, 1998.

[10] Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., and Anthony, M., "Structural Risk Minimization over data-dependent hierarchies." IEEE Trans. Information Theory IEEE Transactions on Vol. 44, Issue 5, pp.:1926-1940, Sept. 1998.

[11] Suarez, A., Saiz-Noeda, M., and Palomar, M., "A Method of Restricted Knowledge Acquisition from WordNet." IEEE Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, pp.38-41 Adelaide, Australia, 1999.

[12] Vapnik, V., "Statistical Learning Theory." Springer, N.Y., 1998.