以標籤區域爲基之網頁文件分類模式

王冠程、黃家偉、呂敏如、林大千、蘇泰郡
南華大學 資訊管理學系

楊士霆
南華大學 資訊管理學系 助理教授
stingyang@mail.nhu.edu.tw

摘 要

隨著網際網路相關技術之盛行，網路使用者亦日趨增加，網路環境資訊量已呈爆炸性成長，因此瀏覽網路上文件或資訊已成爲現代人吸取知識的重要管道之一。故如何有效管理此些網路文件/資訊，讓使用者得以掌握，以協助使用者快速吸收並運用此些網路資訊，乃成爲在現今資訊爆炸時代中之重要課題。目前網頁分類大多以關鍵字擷取或以 HTML 語法標籤內的文字區域爲依據，作爲關鍵資訊分析基礎並進行網頁分類。此些分類技術係將網頁標籤去除，以擷取當中文字型態資訊，進行網頁分類（亦即將所擷取之網頁文字視爲同等重要性），但此種情況下，可能有多項關鍵資訊被忽略（如可能遺失網頁標題資訊）。有鑑於此，本研究提出一套以網頁標籤區域（Tagged-Region）爲基礎之網頁文件分類模式；於模式中，首先本研究乃考量網頁標籤屬性，發展一套「標籤區域權重分配」模組，以尋找影響網頁文件分類之標籤，並解析各網頁標籤於不同網頁空間規劃下之重要性；之後以具分類代表性標籤區域爲基礎，擷取當中關鍵字詞，發展一套「網頁文件類別判定」模組，以推論目標網頁文件之隸屬類別；最後再以鏈結網頁爲基礎，發展一套「鏈結網頁關聯程度推導」模組，將關鍵性鏈結網頁之隸屬類別，修訂目標網頁文件之隸屬類別，以完成網頁文件之隸屬類別判定任務。本研究最終乃建立一套網頁文件自動分類系統，並以一案例評估此模式與技術之有效性與可行性。

綜合言之，本研究之目標乃爲提昇網頁文件分類技術之正確率與效率性，因此，對於資訊需求者而言，本研究期望能協助資訊需求者於龐大之網路資訊/文件中，迅速且便捷地尋得其所需要之網路文件資料，以節省資訊需求者花費於資訊過濾與篩選之大量時間。

關鍵字：標籤區域、網頁文件分類、關鍵字擷取、知識管理

壹、緒論

隨著網際網路相關技術之盛行，網路使用者亦日趨增加，網路環境資訊量已呈爆炸性成長，因此瀏覽網路上文件或資訊已成爲現代人吸取知識的重要管道之一。故如何有效管理此些網路文件/資訊，讓使用者得以掌握，以協助使用者快速吸收並運用此些網路資訊，乃成爲在現今資訊爆炸時代中之重要課題。若能有效地歸類網路資訊或文件，則能有效提高使用者之方便，進而提昇網頁瀏覽率。因網路使用者在進行搜尋資料或學術研究時，往往需要參考並閱讀各種相關資訊，若能使此些資訊有效地符合使用者所需進而歸類，必定能協助使用者更方便且更省時地尋得其所需之資訊，以節省資訊搜尋時間和閱覽不相關之其他資訊。

爲解決上述之問題，目前已發展多種網頁分類技術，如考量網頁內容不僅包含文字，亦包含圖片形式及由數張圖片組合而成之影片形式，因此多數研究乃分析上述兩種資料並進行分類，再將網頁區分等級。此外，亦有研究先將關鍵字儲存於知識庫中，再以此些關鍵字作爲未知類別網頁之分類依據。甚者，由於超文件標示語言主要是以成對出現之標籤來含括某區段之文字（如<title></title>、<h1></h1>等），標籤中所含括之區段文字亦可作爲網頁分類之特徵，相關研究乃利用網頁撰寫者所用的 UML（統一塑模語言）與 HTML（超文件標示語言）之語法，以作爲網頁分類依據。另外，若網頁本身資訊量不足，尚有研究利用網頁內部之超鏈結進行分類（即以超連結之網頁爲基礎，分析連結後網頁之相關資料）。

綜上所述，目前網頁分類大多以關鍵字擷取或以 HTML 語法標籤內的文字區塊爲依據，作爲關鍵資訊分析基礎並進行網頁分類。以上方法雖能將網頁分類，然於資訊量爆增的時代中，未必能將網頁準確地分析和分類、或未必能達到使用者所期望，導致使用者於資訊搜尋效率不佳與耗費時間，其既有之運作模式如圖 1 之 AS-IS Model 所示。

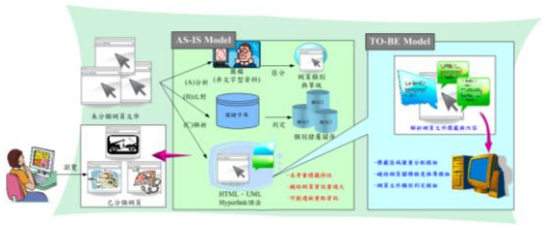

圖 1、網頁文件分類模式之既有與期望模式

如圖 1 所示，目前以超文件標示語言之語法或標籤之分類方式，雖能夠進行網頁分類之任務，讓使用者找尋到所需資料；然而，此種分類法於網頁分類上尚有瑕疵，因此分類技術係將網頁標籤去除，以擷取當中文字型態資訊，進行網頁分類（亦即將所擷取之網頁文字視爲同等重要性），但此種情況下，可能有多項關鍵資訊被忽略，如以主題文字之標籤區域（即含括於<title></title>標籤間之網頁主題文字）爲例，多數主題皆以精簡短句方式呈現，因此主題文字之標籤區域可擷取關鍵字詞較少，或者擷取不到關鍵字詞，然而對於此份網頁文件而言，此標籤區塊之文字資訊應爲重點部分，因此若以此種方式處理，此部分之資訊可能會遺失導致分類效果不佳。此外，若使用鏈結網頁以視爲本身網頁之分析資訊，可能會造成分析資訊量過大，不但分類耗時亦會造成資訊不正確。因此，本研究之研究目的在於先將各網頁標籤進行解析，擷取當中與分類相關之標籤，並探討當中網頁標籤之重要性，並分配對應之權重值，最後利用關鍵字擷取技術以進行網頁文件分類。

綜上所述，目前以超文件語法或標籤分類網頁，本研究乃歸納以下主要之問題：

- 目前多數研究僅分析標籤含括之文字內容，未考量到標籤本身之特性
- 使用鏈結網頁以視爲本身網頁之分析資訊，可能會造成需分析之資訊量過大，分類效率與效果不佳之情況

本研究期望藉由網頁標籤區域之屬性與內容，以進行網頁文件分類任務。因此本研究乃先行分析含括文字型資料之標籤屬性，以分配所有考量標籤之權重值（例如<tltle>所含括字詞較其他標籤具代表性或語意加強標籤<b>等，本研究乃設定較高之權重值）；此外，本研究亦考量相同種類之標籤但位於不同位置標籤區域，其所包含文字區塊重要性亦不盡相同之情形，本研究乃分析各種空間規劃配置下，不同位置標籤區域之重要性分佈狀況。最後爲避免網頁分析資訊量不足或過量之問題，本研究亦以鏈結網頁之類別爲基礎，修訂或微調此目標網頁之隸屬類別，本研究之期望運作模式如圖 1 之 TO-BE Model 所示。

貳、相關文獻探討

本研究所涉及之研究主題乃包括「網頁文件資料解析」及「網頁文件分類」等兩大研究方向，以下即針對此兩項主題之相關研究進行文獻回顧及探討。

(一) 網頁文件資料解析

對於網頁文件之資料解析議題而言，本研究乃針對網頁中標籤資訊、文字型資訊、鏈結資訊及影像等可解析資訊進行相關文獻探討。

(A)網頁標籤資訊

於解析網頁文件內容時，Lim 等人（2005）乃提出以網頁文件中 UML 與 HTML 語法或 Tag 等特徵作爲網頁分類之分析資料，建構一套網頁自動分類系統，該研究乃提出可以擷取網頁中網址與 HTML 語法等特徵，進而提供後續研究之網頁文件分類的分析特徵與資料。此外，宋立群（2006）結合「益助性傾向」和「益助性變化模式」之特質提出一套標籤區域益助性預測機制，此機制可在文件進行分類時，預測未被分析之標籤區域之益助性，進而獲取益助性高的標籤區域以進行分類，當中此標籤區域益助性預測機制乃依分析過之標籤區域之益助性，進行最佳化調整，並且參考相近的益助性特質來預測罕見的視覺形態。

(B)網頁文字型資訊

過去研究常以網頁文字型資訊爲基礎，擷取當中關鍵字，以進行分類。Jenkins

等人（1998）提出 Wolverhampton Web Library 之網頁自動分類計畫，該計畫係利用杜威十進位分類法 DDC（Dewey Decimal Classification），以及人工手動定義關鍵字彙以針對網頁進行分類，雖然該方法利用 DDC 可精準分類網頁，然而尚需動以人力定義關鍵字彙，因此於網頁分類之效率上仍須進行探討與精進。Shen 等人（2007）結合網頁設計與其他數個在 LookSmart 網頁目錄上之最先進摘要演算法，建構一套產生網頁摘要演算法，以提昇網頁分類之效率以及減少多餘之網頁訊息，進而減少使用者於搜尋資料之網頁瀏覽時間。

除上述擷取關鍵字詞外，亦有研究進行關聯語意之解析（Tan 與 Zhang，2008）。Chen 等人（2009）先行針對複合表與關聯表所能引導搜尋之空間範圍及兩者間之搜尋關係，建構語義關連圖 SRG（Semantic Relationship Graph），之後以天真貝式分類器爲主，開發一套以語意關連圖爲基之多關聯天真貝式分類器，該分類器乃根據語意關聯圖之分析結果，排除不必要之特徵與關聯性，進而避免產生無相關之連結。Broder 等人（1997）提出句法相似度（Syntactic Similarity of Files）以解析網頁內文，以判斷各網頁之相似性關係，進而將關聯性較高之兩個網頁予以歸類至同分類中。該研究先將網頁中所有 HTML 標籤去除，再將網頁中各段落內文予以合併，該段落即爲 Shingle，最後藉有兩份網頁之 Shingle 以判斷是否關係。

(C)網頁鏈結資訊

於網頁分類議題中，Furnkranz（2002）乃考量網頁中之超連結網頁資料（Hyperlink Ensembles），以將網頁文件進行分類。該研究乃先從網頁之本文中取得分類特徵，之後考量網頁文件中超連結特徵，連結至對應之網頁，並取得該網頁之所有分類特徵，並且整合與目標網頁之特徵進行整合，以有效地針對目標網頁進行分類。Kuo 與 Wong（2000）提出運用物件轉換模式 OEM（Object Exchange Model）以判定網頁類別，該研究利用 OEM 計算網頁文件之超連結個數，並運用標籤與鏈結內容轉換成 Node Similarity、Edge Similarity 與 Structural Similarity，進而用超鏈結得知整網頁之相似度。另一方面，由於目前的網頁分類技術會因網頁之資訊不足而降低分類準確度，以及因依據過多的鏈結網頁而降低網頁分類之效率，並且受到雜訊的干擾，爲了解決上述問題，許琇娟（2003）建構的分類器係將標籤區域與文件內容分離，該研究首先從文件內容取得關鍵字，並用此關鍵字與每個標籤區域作比對，以獲取標籤區域之關鍵字，之後該研究使用訓練資料所產生的標籤權重，以作爲文件類別的相似度分析，若網頁類別相似度達門檻值，即判定該網頁之類別，反之，則利用鏈結網頁（即尋找與目標網頁具關聯性之網頁內容）之內容解析，以判定目標網頁所屬類別。

(D)網頁影像資訊

網頁內容不僅包含文字，亦包含圖片形式及由數張圖片組合而成之影片形式，故 Fersini 等人（2008）提出分析影像-區塊之技術，以提高網頁分類的準確性，該方法乃分析網頁之影像-區塊，並利用影像-區塊內識別度及資料密集區塊，以確認該影像於網頁中之重要性，並將此網頁特徵納入網頁分類之屬性之一，進而提昇網頁分類準確性。Alpuente 與 Romero（2009）乃以圖像化結構開發一套網頁對照技術。首先，針對 HTML 編碼進行轉譯並擷取圖像化結構網頁中 HTML 標籤，再將網頁標籤轉換時之封包進行壓縮，並依重複性與非重複性以及標籤鏈結的長度是否影響結果，分成平行與垂直兩種結構。分析網頁結構後即得到關於網頁中圖像化之構成要素（即其最小範圍），並以這些範圍條件內尋找關聯網頁，最後將這些關聯網頁以樹狀結構之型式呈現，再以子樹與子樹間之編輯距離來定義網頁間相似度之測量方法，利用網頁之相似度完成分類，不僅可擴大搜尋範圍亦可增加分類效率。

有別於目前文件分類多數分析文字型資料，Wang 等人（2006）乃以每 25 維度爲單位之區域特徵向量，決定圖像檔之各區分區域之區域內容類型。此外，Schettini 等人（2006）建構一套分類與迴歸樹 CART（Classification and Regression Tree）分類器，該分類器乃藉由影像中的低階之感知特徵，以進行數位文件分類。

(二) 網頁文件分類

對於網頁文件分類技術議題而言，過去研究多數以資料探勘技術，以針對網頁所包含之資料特性，進行資料分析與網頁文件分類任務；是故，目前相關文獻應用於網頁文

件之分類技術，則包含結構樹分類法、遺傳演算法、最鄰近區域分類法及類神經網路等分類技術。

(A)結構樹分類法

Wong 與 FU（2000）提出 Labels Discovery Algorithm（LAD），利用 LAD 來獲取網頁完整之階層結構，以正確地區別網頁。該研究利用網頁中標籤建構標籤樹（Tag-Tree），最後利用 Merge Similar Nodeses（即網頁結點）演算法，以獲取網頁完整階層結構，進而使完整結構網頁易於分類。Artail 與 Kassem（2008）乃以網頁中之超文字標記語言 HTML（Hypertext Markup Language）標籤類型之關聯性分析方式，進而縮短 HTML 網頁分類時間與提升其穩定度。該方法架構主要由(1)網頁之清除、(2)頁面之分類並產生子樹、(3)子樹之比較與轉換及(4)分析子樹之相似性等四個階段組成。故該方法論先行從網頁中的中介標籤語言 XML（Meta-Markup Language）檔案擷取資訊，並利用可擴展樣式語言 XSL（Extensible Stylesheet Language）分隔 HTML 檔案與節點以找尋 HTML 之標籤，之後再以 HTML 標籤之特徵分析網頁之關聯性，並運用子樹分類法分類相關聯之網頁，最後再針對網頁內容及標籤之關聯性與相似係數完成分類。

(B)最鄰近區域分類法

Kwon 與 Lee（2003）乃以最鄰近區域分類法 K-NN（K-Nearest Neighbor），協助特徵之選取與標籤權重計算，改善以往文件與文件間之相似特徵分類方法。當中，該方式主要由網頁選擇、網頁分類與網站分類等三個階段組成，故需先行利用全球資源定位器 URL（Universal Resource Locator）找尋網頁搜尋之路徑，再以 BFT（Breath-First Traversal）演算法選擇最短路徑，之後運用 K-NN 演算法針對搜尋後之網頁內容與標籤進行關聯性與相似度分析，並以權重加權方式估計相似度高之內容與標籤，以完成分類任務。Pernkopf（2005）乃以模糊 K 最近點群域 K-NN（K-Nearest Neighbor）分類法輔以遺傳演算法提出一套改善簡易貝式分類器之機制。該研究先行針對搜尋後之網頁內容與標籤進行關聯性與相似度分析，再利用遺傳演算法循序之特徵選擇方式，從特徵子集中選取適合之特徵做爲分類預測屬性，最後再依預測結果完成分類任務。實驗結果顯示，該方法確實改善簡義貝式分類器之準確度。

(C)支援向量機 SVM

爲了有效解決網頁關鍵字分類之同義詞問題，Chen 及 Hsieh（2006）乃提出一個以潛藏語意分析 LSA（Latent Semantic Analysis）與網頁特徵選取 WPFS（Web Page Feature Selection）爲基礎之網頁關鍵資訊選取方法，並結合支持向量機 SVM（Support Vector Machine）之權重投票機制，發展一套網頁分類技術。該研究之細部作法乃以 LAS 技術尋找文件關鍵字與文件之語意關係，並統計各字詞於文件內之隸屬程度，之後以 WPFS 方法萃取網頁文字特徵值，當中，此兩特徵擷取方式係產生不同之結果，因此該研究乃利用 SVM 之權重投票機制，建立關鍵字向量值，最後根據輸出之向量值與投票模式以確定網頁之類別。而研究結果亦顯示該研究能更爲精確地判斷各關鍵字之類別。

(D)貝氏文件分類法

Fujino 等人（2007）針對網頁及科技文件（如學術論文及專利文件等）中多類別與單一標籤之分類領域，提出一套以本文資訊及附屬資訊（如網頁連結、作者文件名稱等資訊）爲分析基礎之整合型文件分類模式。該模式乃利用貝式定理（Bayes）將先行文件內容予以解析，並歸納與分類相關之重要元件（如關鍵字等）及其相對於本文之機率，之後利用多元羅吉斯迴歸模式計算目標文件中各關鍵元件與類別之關係（即各元件之類別隸屬機率值），最後以最大熵值原理獲得此目標文件/網頁與各隸屬類別關係。由於該研究之類別判定模式乃以分析附屬資訊之爲基礎，故相較於其他分類技術而言，該研究更適應於網頁及科技文件之分類中。

此外，Kim 等人（2005）乃結合 Adaptive Boosting 技術與 Uncertainty-based Selective Sampling 技術，提出一套整合性 AdaBUS 技術，以提升貝氏文件分類法 Naïve Bayes Classification（NB）之文件分類準確率。該研究之細部作法乃先將訓練文件進行初步之分類，以獲得初步文件分類結果，之後該研究乃以 Uncertainty-based Selective Sampling（US）技術，尋找分類結果中最不穩定之文件（亦即同時隸屬多個類別之文件），並以人工方式重新分配類別，增加其分類擴增性（Augmentation），同時重新分配各文件類別

及文件屬性之權重。是故，經由數次迭代後所產生之最終分類模式，乃具備文件屬性權重值分配之學習能力，最後研究結果亦顯示，此整合性模式能有效提升以 NB Classification、US 及未修改之 AdaBoost 分類法之文件分類準確率。最後，Youn 與 Jeong（2009）結合天真貝式分類器 NB（Naive Bayes）、特徵比重分類法 CDFW（Class－Dependent－Feature－Weighting）與遞迴特徵消去理論 RFE（Recursive Feature Elimination）建構一套 CDFW－NB－RFE 之文件分類方法。該方法首先從文件中選擇決定性之特徵，並將此些特徵依比重進行分類，再針對分類後之文件特徵進行漸進式篩選，最後將篩選結果作排序，以取得排序爲前之特徵值及其分類屬性，提高分類器之準確率。

(D)關鍵字解析分類

Yang 等人（2000）利用使用者所設定之關鍵字以自動導引至網路並檢索與取得資料，該研究乃採用 Jaccard's Score 以判別網頁相似度，最後顯示網頁之位址（Address）、分數（Score）、抬頭（Title）和符合網頁關鍵字之網頁。此外，多數之文件檢索系統係使用關鍵字以查詢文件，此類系統之作法乃先從文件中擷取文字，之後藉由所建構之權重值分配法則以賦予各關鍵字詞之對應權重值，然此狀況下會產生兩個問題，一爲如何準確的擷取關鍵字，二爲如何確定關鍵字之權重值。有鑑於此，Horng 及 Yeh（2000）提出一套檢索關鍵字方法（稱爲 RK 法），以克服上述之問題。該研究乃使用基因演算法以設定各關鍵字之權重，並且結合 Bigrams（雙字串）、Document Automatic Classification（文件自動分類）、Ranking（排序）和 PAT-tree 模型進行文件關鍵字之檢索，其中任何型態的關鍵字（如人名、地址、技術術語等）皆可被擷取與檢索，藉由上述之研究方法建立與研究實證顯示該研究之分類績效較先前研究爲佳，亦即代表可解決目前之文件關鍵字擷取與權重值設定之問題（Liu 等人，2000）。Chen 等人（2009）提出兩個特徵選取網頁分類技術，該研究乃利用 DPM 減少輸入維度與模糊排序分析之兩階段以分析網頁屬性，進而提昇網頁分類之準確性與效率。

(E)主成分分析

除上述利用資料探勘技術外，亦有研究結合主成分分析方法，以進行網頁分類。Zhang 等人（2009）利用特徵選擇結合 MLNB 提出一套機制，使簡易天真貝式分類器於多元標籤之分類效率上獲得改善。首先，該研究先行利用主要成分分析法（Principal Component Analysis）分析網頁之主要構成要素以擷取特徵，並從特徵集中排除不相關之特徵，之後利用遺傳演算法逐步進化之特性，從特徵子集中選取適合之特徵做爲分類預測屬性，最後再依預測結果完成分類任務。此外，Selamat 與 Omatu（2004）提出新聞網頁分類方法 WPCM（Web Page Classification Method），此方法乃採用類神經網路，先行取得主要成分和使用者導向（Profile-based）之分類特徵。當中新聞網頁係由詞彙加權（Term Weighting）方案所擷取，然而需蒐集相當大量之詞彙，因此乃使用主成分分析（PCA）取得高度相關之特徵，PCA 之結果乃包含各類別中最普遍之詞之階級輪廓（Class-Profile）結合特徵向量。手動地選擇自各類別之普遍詞，並自加權部分使用熵權重（Entropy Weighting）方案，自各類別之固定數量之普遍詞中使用特徵向量，亦自 PCA 中減少主要之成分，最後，將此特徵向量輸入至神經網絡進行分類，已達到分類之準確度。

(F)其他分類技術

Chen 等人（2006）結合「公平特徵集合選取」FFSS（Fair Feature-Subset Selection）演算法和「適應性模糊學習網路」AFLN（Adaptive Fuzzy Learning Network）演算法，提出一套智慧型網頁文件自動分類模式。首先，FFSS 演算法乃公正地選取並處理各類別之特徵，並辨識得當中具顯著分類之特徵，進而縮小特徵選取之範圍；其次，AFLN 乃提供快速之學習能力模型，該演算法可藉由不斷的系統訓練，自動地糾正不明確的分類行爲，並藉由上述兩個演算法之整合，即可更有效地改善網頁文件分類績效。

此外，Jenkins 與 Inman（2000）提出可調適自動化之網頁分類模式與技術，該模式乃分析訓練網頁中出現頻率較高之字彙與 HTML 標籤屬性，並利用字彙自動產生分類時所需使用之分類字彙，進而產生階層式之分類節點，最後根據階層式之分類節點上的分類字彙，以針對測試網頁訂定類別；甚者，該研究亦可依據不同類型文件所使用之語

言，自動調整產生分類字彙。此外，Lin 等人（2002）建構一套 ACIRD（Automatic Classifier for the Internet Resource Discovery）智慧型文件分類與檢索系統，該系統主要乃包含文件知識擷取機制、文件分類器與兩階段式搜尋引擎三部份，利用此三部份以提升網路文件之分類處理效率，當中該系統係利用知識擷取機制，針對網路上已分類之文件進行知識擷取與吸收，並利用文件分類所學習之知識（即文件屬性），針對新進文件進行分類，最後使用者可透過系統之兩階段搜尋引擎，搜尋得所所欲之知識文件。

此外，機器翻譯係自然語言處理研究上重要課題之一。過去運用機器翻譯較成功之例子，多爲特定領域文件之翻譯。然由於網際網路與搜尋引擎之盛行，機器翻譯在跨語言檢索（Cross-Language Information Retrieval）中的角色開始受到重視（Oard 與 Resnik，1999）。

除上述分類技術外，於網頁分群議題中，Runkler 及 Bezdek（2003）提出利用校正距離（即辨識字串相似度距離；Levenshtein Distance）及圖示距離（Graph Distance），將非數值資料轉換爲關聯資料集以進行分析（如網頁內容與瀏覽網頁紀錄等資料型態），之後透過 RACE 模式（Relational Alternating Cluster Estimation）進行分群與相關分析（亦即利用關聯性資料相對關係作爲相似度距離，以推論得分群）。當中，文字型態資料係以關鍵字分析技術以達文件分類與自動歸檔之目的，此外瀏覽網頁紀錄則可用以分析使用者偏好，以作爲區隔使用者偏好、網站內容與網站類別之參考依據。

參、系統功能簡介

整合不同使用者所需之功能，本系統開發之重點共可分爲「網頁文件維護模組」、「網頁文件類別判定模組」、「關鍵字維護模組」、「系統權重模組」及「使用者資料維護模組」等五大模組，系統乃依據使用者之權限開放對應之系統功能供其使用，圖 2 即爲此網頁文件分類模組功能架構。

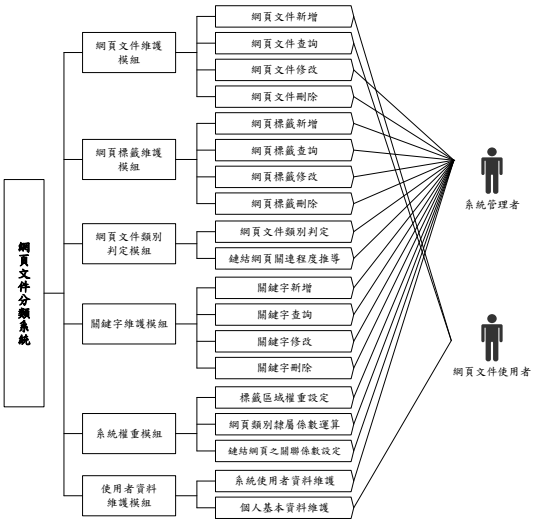

圖 2、網頁文件類別判定模組功能架構

各模組之性能說明如下：

（1）網頁文件維護模組

- 網頁文件新增：可上傳欲分類之網頁文件，並將資料進行擷取
- 網頁文件查詢：可查詢所上傳之網頁文件資料或已分類之網頁文件
- 網頁文件修改：網頁文件使用者可將已分類過後之網頁文件下載
- 網頁文件刪除：可將已上傳之網頁文件刪除

（2）網頁文件類別判定模組

- 網頁空間解析：可對已擷取之網頁文件資料進行空間配置解析，衡量各個空間位置不同之重要性
- 標籤區域權重分配：可依據不同重要性之網頁空間內之標籤區域進行不同權重分配
- 文件類別判定：將已上傳之網頁文件進行類別判定
- 鏈結網頁關聯程度推導：可將已擷取之鏈結網頁和目標網頁進行比對，分析其關聯程度

（3）關鍵字維護模組

- 關鍵字庫維護：可新增、查詢、修改、刪除關鍵字庫內之資料

（4）網頁標籤維護模組

- 網頁標籤維護：可新增、查詢、修改、

刪除網頁標籤內之資料

（5）系統權重模組

- 標籤區域權重設定：可依據不同重要性之標籤區域進行其權重設定
- 網頁類別隸屬係數運算：可進行標籤區域內之關鍵字所應隸屬類別之係數運算
- 鏈結網頁關聯係數設定：可將已擷取之鏈結網頁和目標網頁進行比對，設定其中之關聯係數
- 類別係數之正規化：可依照權重之設定及判定後類別係數之比對，將各類別計算過之係數進行修正

（5）使用者資料維護模組

- 系統使用者資料維護：系統管理者可進行新增、查詢、修改、刪除使用者之基本資料
- 個人基本資料維護：可提供網路使用者新增、查詢、修改、刪除其基本資料

肆、系統特色

整合不同使用者所需之功能，本系統開發之重點共可分爲「網頁文件維護模組」、「網頁文件類別判定模組」、「關鍵字維護模組」、「系統權重模組」及「使用者資料維護模組」等五大模組，，各模組之性能說明如下：

（1）網頁文件維護模組

- 網頁文件新增：可上傳欲分類之網頁文件，並將資料進行擷取
- 網頁文件查詢：可查詢所上傳之網頁文件資料或已分類之網頁文件
- 網頁文件修改：網頁文件使用者可將已分類過後之網頁文件下載
- 網頁文件刪除：可將已上傳之網頁文件刪除

（2）網頁文件類別判定模組

- 網頁空間解析：可對已擷取之網頁文件資料進行空間配置解析，衡量各個空間位置不同之重要性
- 標籤區域權重分配：可依據不同重要性之網頁空間內之標籤區域進行不同權重分配
- 文件類別判定：將已上傳之網頁文件進行類別判定
- 鏈結網頁關聯程度推導：可將已擷取之鏈結網頁和目標網頁進行比對，分析其關聯程度

（3）關鍵字維護模組

- 關鍵字庫維護：可新增、查詢、修改、刪除關鍵字庫內之資料

（4）網頁標籤維護模組

- 網頁標籤維護：可新增、查詢、修改、刪除網頁標籤內之資料

（5）系統權重模組

- 標籤區域權重設定：可依據不同重要性之標籤區域進行其權重設定
- 網頁類別隸屬係數運算：可進行標籤區域內之關鍵字所應隸屬類別之係數運算
- 鏈結網頁關聯係數設定：可將已擷取之鏈結網頁和目標網頁進行比對，設定其中之關聯係數
- 類別係數之正規化：可依照權重之設定及判定後類別係數之比對，將各類別計算過之係數進行修正

（5）使用者資料維護模組

- 系統使用者資料維護：系統管理者可進行新增、查詢、修改、刪除使用者之基本資料
- 個人基本資料維護：可提供網路使用者新增、查詢、修改、刪除其基本資料

伍、研究方法

本系統乃建置於 Microsoft WinXP 作業系統上，以 JSP 語言開發系統之各項功能，採用 Microsoft SQL Server 2005 資料庫系統儲存系統運作過程之相關資料；以下乃分別介紹系統開發時所使用之工具。

JSP（Java Server Pages）

JSP 是由 Sun Microsystem 公司倡導，其並集合其他公司共同建立之動態網頁技術標準。由於 JSP 乃是以 Java 程式語言爲基礎之網站伺服器描述語言程式，故其乃繼承 Java 支援跨平台與跨網站伺服器之優點，使網頁

設計更具彈性。當使用者透過瀏覽器向伺服器端要求開啓 JSP 網頁時，架設於伺服器端上之 JSP 引擎乃先將 JSP 網頁轉譯爲 Servlet，其次再將 JSP 執行後所產生之 HTML 文件傳送至用戶端，並同時顯示執行結果於瀏覽器上。因此，用戶端於瀏覽器中所見之內容並非 JSP 網頁之原始內容，而是 JSP 網頁執行後所產生之 HTML 文件。此外，JSP 尚具有以下特性：

- 瀏覽者端環境：各種網頁瀏覽器均可，如 MSIE、KKMan 等。
- 伺服器端環境：Windows NT/2000、Linux、Unix 和 Mac 等，並加上「J2SDK」Java 程式編譯工具與 Tomcat 等 JSP 伺服器。
- 伺服器端搭配資料庫：如 SQL Server、MySQL、Oracle 等資料庫系統。
- 平台與伺服器之獨立性：JSP 技術一次寫入之後，可以在任何具有符合 JavaTM 語法結構的環境下執行。
- 模組程式之可重用性：JSP 元件（Enterprise Javabeans、Javabeans 或自訂之 JSP 標籤）皆爲跨平台且可重用之元件。而 Enterprise JavaBeans 元件可存取傳統資料庫，並可以分散式系統模式於 Unix 與 Windows 平台工作，減少程式開發時間、增加程式彈性。
- 執行效率佳：JSP 僅於第一次執行時被編譯成 Java Servelt，並載入記憶體中以便下次瀏覽，故除非網頁更新，否則系統無需重新編譯。
- 與 HTML 緊密整合：由於 JSP 支援伺服端 Scripting 語言之環境，因此 JSP 可嵌入 HTML 標籤中使用，不僅提高便利性亦減少 I/O 問題。
- 標籤可擴充性：由於 JSP 技術兼容 XML 標籤技術，使程式開發者能自訂標籤庫，並充分利用與 XML 相容之標籤技術功能，減少對 Scripting 語言之依賴，降低網頁製作者於製作網頁與擴充網頁功能之複雜度。

關聯式資料庫—Microsoft SQL Sever 2005

Microsoft SQL Server 2005 爲一關聯式資料庫（Relational Database Management Systems），此種資料庫採資料分類表格化之架構，將相關資料組成表格，且表格間具有關聯性。其優點在於其所含之各資料表可獨立運作，修改資料表內容時不會互相影響，且查詢時可藉由各資料表間之關聯性；其可利用 SQL 語法進行資料查詢，以快速擷取所需之資料。此外，由於 Microsoft SQL Server 2005 具有與網際網路應用程式相容之特性，且 SQL 語法可配合各種程式如 VB（進行本機資料庫處理）、JSP（進行遠端資料庫處理）等進行大量資料之處理與運算，故使用 Microsoft SQL Server 2005 作爲系統後端資料庫，可方便維護資料庫之資料結構、查詢、新增、修改或刪除資料表之內容。

陸、系統使用對象

爲提升網路使用者查詢網頁之效率，本研究乃發展一套以標籤區域爲基之網頁文件分類系統，使用者可藉由輸入欲查詢之網頁文件名稱或將網頁文件上傳至系統，進而使用本系統所提供之各項功能。爲使系統運作順暢，本系統乃將使用者分爲網路查詢使用者、系統管理者二種不同角色。以下即分別針對此二類型之使用者於本系統中可使用之功能進行描述：

網路查詢使用者

1. 可輸入欲查詢之網頁文件名稱
2. 可上傳網頁檔案進行分類
3. 可查閱/下載相同分類之網頁文件
4. 可查詢、修改、刪除個人基本資料

系統管理者

1. 可新增、查詢、修改、刪除所擷取之網頁文件資料
2. 可新增、查詢、修改、刪除關鍵字庫

3. 可新增、查詢、修改、刪除
4. 可執行網頁空間規劃之解析、配置
5. 可新增、查詢、修改、刪除所擷取之文字型標籤資料
6. 可進行網頁文件類別之判定
7. 可修改、查詢系統權重值
8. 可新增、查詢、修改、刪除所有系統使用者之基本資料

整合不同使用者所需之功能，本系統開發之重點共可分爲「網頁文件維護模組」、「網頁文件類別判定模組」、「關鍵字維護模組」、「系統權重模組」及「使用者資料維護模組」等五大模組，系統乃依據使用者之權限開放對應之系統功能供其使用，圖 3 即爲此網頁文件分類模組功能架構。

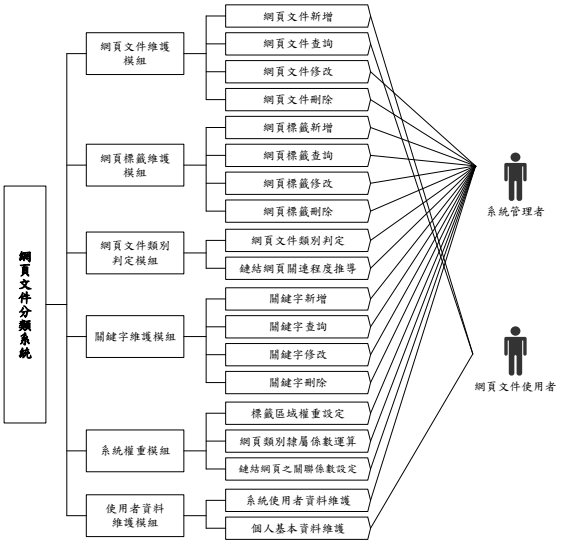

圖 3、網頁文件類別判定模組功能架構

柒、系統使用環境

系統管理者將系統權重設定完畢後，系統即開始運作。由於本研究是以「google 新聞、yahoo 奇摩新聞」爲驗證案例，因此網頁文件使用者可先行於 google 新聞、yahoo 奇摩新聞或聯合新聞網等網站下載欲進行分類之網頁文件，下載時，將檔案存成 HTML 檔或是 MHT 檔，如圖 4 所示，以下載「疑似飛碟在南華大學上空出沒」之網路新聞爲案例，下載完畢後，即可利用本系統進行網頁文件分類。

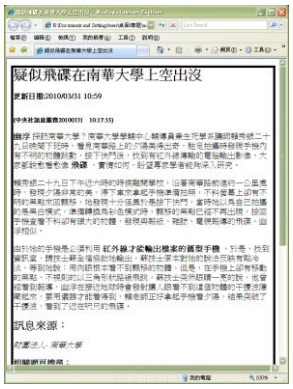

圖 4、網路新聞「疑似飛碟在南華大學上空出沒」

待網頁文件使用者將資料蒐集完畢後，網頁文件使用者即可將欲判定類別之網頁文件，透過「網頁文件上傳功能」上傳至系統中，以使系統判定目標網頁文件之類別。因此如圖 5 所示，「網頁文件上傳功能」乃提供網頁文件使用者將網頁文件基本資料匯入系統資料庫內，例如網頁文件使用者依序輸入此網頁文件名稱爲「疑似飛碟在南華大學上空出沒.htm」、類別爲「資訊」、關鍵字爲「南華大學」、「南華」、「飛碟」與摘要等網頁基本資料，並瀏覽上傳網頁檔名爲「疑似飛碟在南華大學上空出沒.htm」之檔案（詳見圖 5 之內容）；最後，網頁文件使用者按下「確定」鍵後，即完成網頁上傳作業，且系統同時擷取網頁文件中網頁頭部標籤、網頁主體標籤、超連結標籤等區域之網頁資料，以作爲網頁文件分類之分析基礎。

圖 5、「網頁文件維護模組」上傳功能

當網頁文件使用者將網頁文件蒐集並上傳至系統完畢後，即交由系統管理者進行網頁文件類別判定之動作。系統管理者可利用「網頁文件類別判定模組」來完成動作。在「網頁文件類別判定模組」中首先界定網頁標籤區域，再擷取目標網頁文件中各網頁標籤區域所包含之關鍵字，利用目標網頁文件關鍵字出現頻率、領域關鍵字與類別關係之訓練資料庫，以及參照標籤區域權重分配模組所建構之標籤區域權重分配表，進而計算目標網頁文件與各類別之關係係數，以初步判定此目標網頁文件類別偏向，但目標網頁文件之關係係數總和不為 1，則將目標網頁文件與類別之關係係數予以正規化，可得另一係數（即目標網頁文件與類別之類別隸屬係數），即可獲知目標網頁文件之隸屬類別。

當系統管理者執行「網頁文件類別判定」功能時，系統乃提供查詢欄位供系統管理者輸入，輸入完畢後即會出現符合查詢條件之網頁文件資料，系統管理者即可根據資料勾選欲分類之網頁文件。若系統管理者勾選「大學加 X 業者不推、學生興趣缺」之網頁文件，並按下「資料送出」(如圖 6 所示)，經過系統計算後即可得知目標網頁文件之類別為，「教育」、「社會」、「生活」、「科技」、「藝文」之類別隸屬係數為「0.7」、「0.28」、「0.01」、「0」、「0」，因「教育」類別之類別隸屬係數為「0.7」，大於「網頁文件類別判定門檻值」（系統預設為 0.5），越大代表越趨近該類別，系統即說明此目標網頁文件較為趨近於「教育」類別（如圖 7 所示）。

圖 6、網頁文件類別判定模組

圖 7、目標網頁文件類別判定（隸屬係數大於 0.5）

若系統管理者勾選「疑似飛碟在南華大學上空出沒」之網頁文件，(如圖 6 所示)，經過系統計算後即可得知目標網頁文件之類別為「教育」、「社會」、「生活」、「科技」、「藝文」與此類別隸屬係數為「0.47」、「0.35」、「0.11」、「0.07」、「0」(如圖 8 所示)，由於此網頁類別隸屬係數皆小於「網頁文件類別判定門檻值」(系統預設為 0.5)，則系統會建議使用者繼續進行鏈結網頁關聯程度推導。

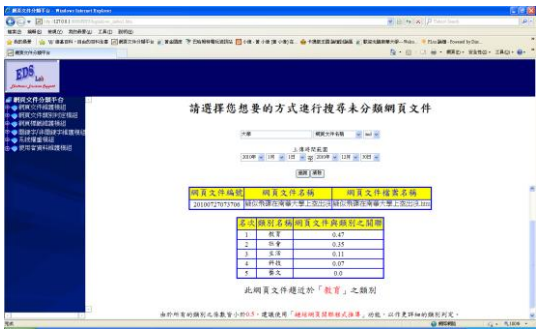

圖 8、目標網頁文件類別判定（隸屬係數小於 0.5）

若網頁文件使用者採納系統建議，欲將網頁類別隸屬係數小於「網頁文件類別判定門檻值」（系統預設爲 0.5）之網頁文件做更精確之判定，則交由系統管理者進行「鏈結網頁關聯程度推導」動作，乃利用「鏈結網頁關聯程度推導」模組，以歸納與目標網頁高度相關之鏈結網頁，進而修正目標網頁之隸屬類別。

當系統管理者執行「鏈結網頁關聯程度推導」功能時，系統乃提供查詢欄位供系統管理者輸入。當搜尋條件組合完畢且送出後，系統即可取得網頁文件名稱「疑似飛碟在南華大學上空出沒」與「台灣大學未來競爭力 國際視野是關鍵」之二筆網頁文件資料（如圖 9 所示），之後使用者可勾選需要修訂類別之目標網頁文件，資料送出後，系統乃先行擷取此目標網頁所有對應之鏈結網頁，再者系統乃計算各網頁之鏈結關聯係數，並參照鏈結網頁預選等級（如圖 10 所示），最後系統乃選定排序前「三」名鏈結網頁，並賦予對應之權重值（如圖 10 所示），進而修訂目標網頁文件之所屬類別。

因此，經由「鏈結網頁關聯程度推導」模組運算後，即可得知目標網頁文件檔案名稱「疑似飛碟在南華大學上空出沒.htm」之類別爲「科技」、「社會」、「教育」、「生活」與各類別之隸屬係數爲「0.55」、「0.25」、「0.16」、「0.04」，系統並提供網頁文件類別判定之類別與係數給予使用者對照，經由對照之後發現此目標網頁文件由原本網頁文件類別判定爲「教育」類別之係數「0.47」，經鏈結網頁關聯程度推導修正後，該網頁文件與「科技」類別之係數亦修正爲「0.55」（如圖 11 所示）。

圖 9、鏈結網頁關聯程度推導

圖 10、鏈結網頁預選等級

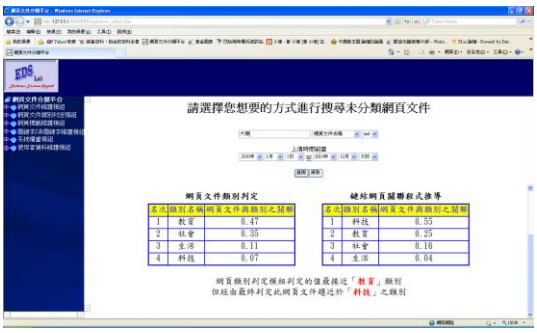

圖 12、鏈結網頁關聯程度推導

參考文獻

1. 宋立群，2006，「漸進式網頁文件分類技術」，博士論文（指導教授：郭經華），淡江大學資訊工程學系博士班。
2. 孫銘聰，2002，「啓發式電子化文件權限推論模式與技術構建」，碩士論文（指導教授：侯建良），國立清華大學工業

工程與工程管理學系。
3. 許琇娟，2000，「以漸進式標籤區域分析爲基礎之網頁分類器」，碩士論文（指導教授：林丕靜），淡江大學資訊工程學系。
4. Alpuente, M. and Romero, D., 2009, "A Visual Technique for Web Pages Comparison," *Electronic Notes in Theoretical Computer Science*, Vol. 235, No. 1, pp. 3-18.
5. Artail, H., Kassem, F., 2008, "A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations," *Data & Knowledge Engineering*, Vol. 66, No. 2, pp. 326-337.
6. Broder, A. Z., Glassman, S. C., Manasse, M. S. and Zweig, G., 1997 "Syntactic clustering of the Web," *In Proceedings of the Sixth International World Wide Web Conference, Santa Clara, California USA, April 7-11*, pp. 391-404.
7. Chen, C. M., Lee, H. M. and Chang, Y. J., 2009, "Two novel feature selection approaches for web page classification," *Expert Systems with Applications*, Vol. 36, No. 1, pp. 206-272.
8. Chen, C. M., Lee, H. M. and Tan, C. C., 2006, "An intelligent web-page classifier with fair feature-subset selection," *Engineering Applications of Artificial Intelligence*, Vol. 19, No. 8, pp. 967-978.
9. Chen, H., Liu, H., Han, J., Yin, X. and He, J., 2009, "Exploring optimization of semantic relationship graph for multi-relational Bayesian classification," *Decision Support Systems*, Vol. 48, No. 1, pp. 112-121.
10. Chen, R. C. and Hsieh, C. H., 2006, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Systems with Applications*, Vol. 31, pp. 427-435.
11. Chen, R. C. and Hsieh, C. H., 2006, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Systems with Applications*, Vol. 31, No. 2, pp. 427-435.
12. Fersini, E., Messina, E. and Archetti, F., 2008, "Enhancing web page classification through image-block importance analysis," *Information Processing & Management*, Vol. 44, No. 4, pp. 1431-1447.
13. Fujino, A., Ueda, N. and Saito K., 2007, "A hybrid generative/discriminative approach to text classification with additional information," *Information Processing and Management*, Vol. 43, pp. 379-392.
14. Furnkranz, J., 2002, "Hyperlink ensembles: A case study in hypertext classification," *Information Fusion*, Vol. 3, No. 4, pp. 299-312.
15. Horng, J. T. and Yeh, C. C., 2000, "Applying genetic algorithms to query optimization in document retrieval," *Information Processing and Management*, Vol. 36, pp. 737-759.
16. Hou, J. L. and Lin, F. H., 2004, "A document and user matching model via document keyword analysis," *Journal of Computer Information Systems*, Vol. 44, No. 4, pp. 1-15.
17. Jenkins, C. and Inman, D., 2000, "Adaptive automatic classification on the Web," *In proceedings of the 11th international workshop, Database and Expert Systems Applications*, pp 504-511.
18. Jenkins, C., Jackson, M., Burden, P. and Wallis, J., 1998, "Automatic classification of Web resources using Java and Dewey Decimal Classification," *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 646-648.
19. Kim, H. J., Kim, J. U. and Ra, Y. G.., 2005, Boosting Naïve Bayes text classification using uncertainty-based selective sampling," *Neurocomputing*, Vol. 67, pp. 403-410.
20. Kuo, Y. H., Wong, M. H., 2000, "Web document classification based on hyperlinks and document semantics," *PRICAI 2000 Workshop on Text and Web Mining,*

Melbourne, Australia August 2000, pp. 44-51.

21. Kwon, O. W., Lee, J. H., 2003, "Text categorization based on k-nearest neighbor approach for Web site classification," *Information Processing & Management*, Vol. 39, No. 1, pp. 25-44.
22. Lim, C. S., Lee, K. J. and Kim, G. C., 2005, "Multiple sets of features for automatic genre classification of web documents," *Information Processing & Management*, Vol. 41, No. 5, pp. 1263-1276.
23. Lin, S. H., Chen, M. C., Ho, J. M. and Huang, Y. M., 2002, "ACIRD: Intelligent Internet document organization and retrieval," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 3, pp. 599-614.
24. Liu, C. H., Lu, C. C. and Lee, W.-P., 2000, "Document categorization by genetic algorithms," *IEEE International Conference on Systems*, Vol. 5, pp. 3868-3872.
25. Oard, D. W. and Resnik, P., 1999, "Support for interactive document selection in cross-language information retrieval," *Information Processing and Management*, Vol. 35, pp. 363-379.
26. Pernkopf, F., 2005, "Bayesian network classifiers versus selective k-NN classifier," *Pattern Recognition*, Vol. 38, No. 1, pp. 1-10.
27. Runkler, T. A. and Bezdek, J. C., 2003, "Web mining with relational clustering," *International Journal of Approximate Reasoning*, Vol. 32, No. 2, pp. 217-236.
28. Schettini, R., Brambilla, C., Ciocca, G., Valsasna, A. and Ponti, M. D., 2002, "A hierarchical classification strategy for digital documents," *Pattern Recognition*, Vol. 35, No. 8, pp. 1759-1769.
29. Selamat, A. and Omatu, S., 2004, "Web page feature selection and classification using neural networks," *Information Sciences*, Vol. 158, pp. 69-88.
30. Shen, D., Yang, Q. and Chen, Z., 2007, "Noise reduction through summarization for Web-page classification," *Information Processing & Management*, Vol. 43, No. 6, pp. 1735-1747.
31. Tan, S. and Zhang, J. 2008, "An empirical study of sentiment analysis for Chinese documents," *Expert Systems with Applications*, Vol. 34, pp. 2622 – 2629.
32. Wang, Y., Phillips, I. T., and Haralick, R. M., 2006, "Document zone content classification and its performance evaluation," *Pattern Recognition,* Vol. 39, No. 1, pp. 57-73.
33. Wong, W. C., and Fu, W. C A., 2000, "Finding Structures of Web Documents," *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Dallas, TX., USA, May 14.
34. Yang, C. C., Yen, J. and Chen, H., 2000, "Intelligent internet searching agent based on hybrid simulated annealing," *ELSEVIER Journal on Decision Support System*, pp. 269-277.
35. Youn, E. and Jeong, M. K., 2009, "Class dependent feature scaling method using naive Bayes classifier for text datamining," *Pattern Recognition Letters*, Vol. 30, No. 5, pp. 477-485.
36. Zhang, M. L., Peña, J. M., Robles, V., 2009, "Feature selection for multi-label naive Bayes classification," *Information Sciences*, Vol. 179, No. 19, pp. 3218-3229.