行政院國家科學委員會專題研究計畫 成果報告

以標籤區域為基之網頁文件分類模式
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2221-E-343-004-
執 行 期 間 ：99 年 08 月 01 日至 100 年 07 月 31 日
執 行 單 位 ：南華大學資訊管理學系

計 畫 主 持 人：楊士霆

計畫參與人員：碩士班研究生-兼任助理人員：龔鈺婷
大專生-兼任助理人員：黃家偉

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 100 年 08 月 09 日

壹、中英文摘要

中文摘要

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為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，乃成為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為依據，作為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為同等重要性），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）為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為基礎，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；最後再以鏈結網頁為基礎，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，本研究之目標乃為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

Abstract

Owing to the booming growth of Internet technology, the number of Web documents has significantly increased over the Internet. If the Web documents can be effectively managed, the knowledge demanders (i.e., Internet users) can efficiently absorb and use the knowledge documents; it has become the core topic in this information explosion era. Web document classification technology with high accuracy can improve the efficiency for Internet users to search required knowledge and to save lots of knowledge-searching time. Concerning complexity of Web page structure, this paper analyzes the tagged-region characteristics including tag attributes and tag locations of Web page to develop an algorithm for web document classification. Based on tagged-region characteristic analysis, each tag can be identified and given different weighting value. Therefore, the keywords extracted from each tagged-region are weighted and then the categories of the target Web document can be determined. Furthermore, based on the hyperlink tag, the similar Web documents can be collected to re-determine target Web document categories.

In addition to the Web document classification algorithm, a Web-based Web document classification system is also developed and a demonstration case is applied to verify the performance of the proposed approach. The attempt of this research is to enhance the accuracy and efficiency of Web document classification technology and to enable a Web knowledge management mechanism over the Internet.

Keywords: Tagged-Region, Web Document Classification, Keyword Extraction, Knowledge Management

一、研究動機與目的

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為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，乃成為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。

為解決上述之問題，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，再以此些關鍵字作為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為網頁分類之特徵，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，以作為網頁分類依據。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（即以超連結之網頁為基礎，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為依據，作為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。

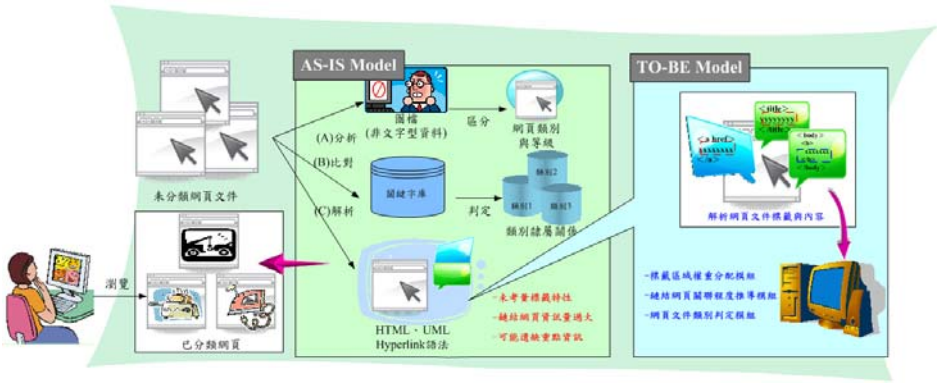

圖 1、網頁文件分類模式之既有與期望模式

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為同等重要性），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）為例，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為重點部分，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為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
- 使用鏈結網頁以視為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。最後為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為基礎，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。

2. 文獻回顧

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。

2.1 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 等特徵作為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為基礎，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為主，開發一套以語意關連圖為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，該段落即為 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，為了解決上述問題，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，以作為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 維度為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。

2.2 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為分類預測屬性，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

為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）為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為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）為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為基礎，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，以取得排序為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，一為如何準確的擷取關鍵字，二為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（稱為 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為佳，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為分類預測屬性，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為，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，多為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），將非數值資料轉換為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為相似度距離，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，以作為區隔使用者偏好、網站內容與網站類別之參考依據。

3. 以標籤區域為基之網頁文件分類模式

本研究所提出之「以標籤區域為基之網頁文件分類模式」乃以網頁標籤區域（Tagged-Region，亦即被成對標籤分割而成的文字區域）為分析基礎，先行尋找各網頁中含括文字型資料之標籤並擷取當中關鍵字詞，之後配合網頁空間規劃與網頁標籤之屬性，針對各標籤區域設定對應之權重值（如<tltle>所含括字詞較其他標籤具代表性，或語意加強標籤<b>（粗體標籤）等標籤，本研究乃設定其所對應之權重值較高），之後即可依照各標籤區域所擷取關鍵字詞及標籤區域權重分配，並配合領域關鍵字與類別之隸屬關係（需由領域專家先行建置），即可得知此目標網頁文件之所屬類別，最後，利用網頁鏈結標籤（即<a href>）尋找當中具高度關聯性之鏈結網頁，並依其類別隸屬關係，以修訂目標網頁文件之類別。因此本研究之主要流程可分為三大部份，分別為「標籤區域權重分配模組」（如圖 2 之 Part1 所示）、「網頁文件類別判定模組」（如圖 2 之 Part2 所示）及「鏈結網頁關聯程度推導模組」（如圖 2 之 Part3 所示）。

3.1 標籤區域權重分配模組

於網頁文章之設計過程中，多數網頁設計者係利用超文件標示語言（Hyper Text Markup Language；HTML）以進行文章內容之撰寫與編排。由於超文件標示語言，主要係利用成對出現的標籤以指定網頁文字之呈現方式，是故，網頁設計者僅需運用適當之網頁標籤，網頁文件即可設計與傳統文章相同之閱讀模式。如同一般文章之寫作方式，網頁設計者即可利用網頁標籤以設定網頁文件之標題、摘要、關鍵字、章節標題等內容，此外，網頁設計者亦常以粗體、斜體或加底線之方式強調字詞之獨特性與重要性。其次，除上述標籤屬性各具其代表性外，相同種類之標籤但位於不同位置標籤區域（即外部標籤亦可含括內部標籤），其所包含文字區塊之重要性亦不盡相同，因此為了區別不同種類與相同種類但不同位置的標籤區域，本研究乃解析網頁空間規劃（即標籤區域配置關係），以進行各空間規劃配置下之標籤區域權重設定。

綜上所述，於網頁標籤區域權重分配模組中，本研究乃先擷取網頁文件中與分類相關（即含括

文字型資料）之標籤（即分類依據標籤擷取機制），之後針對不同空間規劃下標籤區域權重分配（即標籤位置解析機制）進行探討，分別如以下說明之。

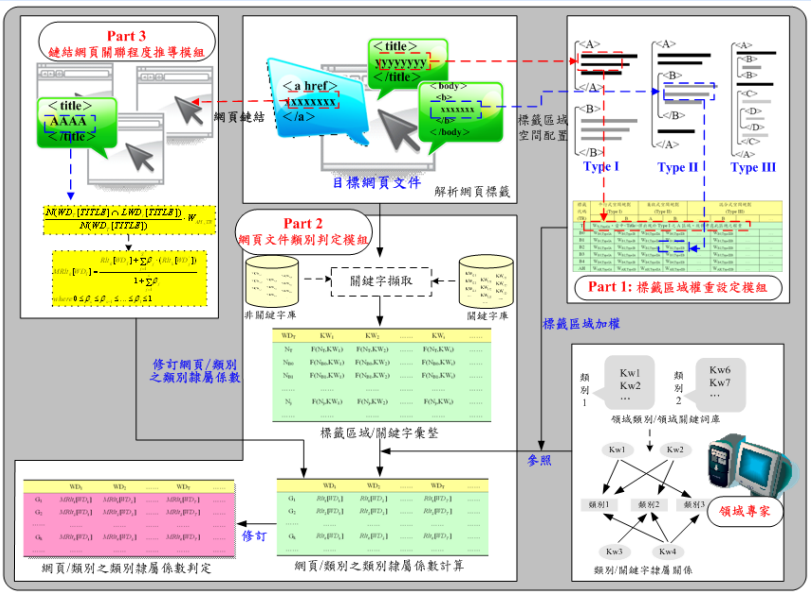

圖 2、以標籤區域為基之網頁文件分類方法論之流程架構

3.1.1 分類依據標籤擷取機制

於網頁文件標示語言中，所有的標籤（包含文件類型聲明標籤、HTML 文件宣告標籤、網頁頭部標籤、網頁主體標籤、註解標籤等）皆具備其特定之用途，而被成對標籤所含括的文字內容，通常亦反映出其所屬標籤的性質，如「網頁頭部標籤」（包含<head>、<title>、<bgsound>、<meta>、<style>及<script>等標籤）主要用以宣告整個文件相關格式、型態、網頁名稱、Script 描述語言及樣式表等設定值之區段，當中，標題文字標籤<title>除代表此網頁之名稱外，設計者亦常利用標籤<h1>至標籤<h6>（即顯示 6 種大小不同標題文字），以突顯出主題的字彙。

此外，「網頁主體標籤」（即<body>）乃含括所有欲顯示於網頁之文字、圖形及其它的多媒體文件等（即需置於<body>與</body>標籤之中），當中於網頁文字顯示方面，網頁設計師常使用粗/斜體字標籤（包含<b>、、<cite>及等標籤）、「列表法標籤」（包含<ol>、<li>、<ul>、<dl>、<dt>及<dd>等）及「引述文字標籤」（<blockquote>）以強調網頁中所顯示字詞之重要性。

綜上所述，於一份網頁文件中，被不同之標籤所包含之文字區域，其所代表之意義與重要性亦尚存差異，因此，由於本研究之網頁分類模組係以網頁中文字型資料為分析依據，亦即以網頁標籤中「網頁頭部標籤」及「網頁主體標籤」為主，並考量當中文字語調加強標籤，以作為本研究分類依據之標籤，當中所需探討之網頁標籤彙整如表 1。

表 1、分類依據標籤列表

標籤類型	標籤名稱	重點文字標籤	
網頁頭部標籤	<title>	標題文字設定標籤 (T)	<h1>、<h2>、<h3>、<h4>、<h5>、<h6>
網頁主體標籤	<body>	粗體文字標籤 (B1)	<b>、
		斜體文字標籤 (B2)	<address>、<cite>、<dfn>、、<i>
		列表法文字標籤 (B3)	、 、 、<dl>、<dt>、<dd>
		引述文字標籤 (B4)	<blockquote>
超連結標籤	<a href >	導引網路節點標籤 (AH)	

透過上述之分類依據標籤擷取機制，本研究乃於網頁文件進行分類時，首先儲存各標籤中文字區域之內文，並擷取當中具代表性之關鍵字，之後考量各網頁標籤所強調重要性進而賦予不同權重值（如標題文字標籤或粗/斜體字標籤等），以進行網頁類別隸屬係數之計算與類別之判定。

3.1.2 標籤位置解析機制

待上述分類依據標籤擷取（即文字型標籤）完成後，由於相同種類之標籤但位於不同位置之標籤區域，其所包含文字區塊之重要性亦不盡相同，為了區別不同種類與相同種類但不同位置的標籤區域，因此本研究乃參考許琇娟（2000）之網頁空間規劃，以進行位處不同位置之標籤區域權重設

定。

由於超文件標示語言具有網頁文件空間規劃的功能，因此本研究須進一步探討標籤區域於網頁空間上關係。許琇娟（2000）指出常見之網頁文件空間規劃係劃分為三種：(1)平行式空間規劃、(2)巢狀式空間規劃及(3)混合式空間規劃（本研究乃分別將其定義為 TypeI、TypeII 及 TypeIII，如圖 3 至圖 5 所示），首先將標籤位置解析機制所使用之符號定義，並說明各空間規劃之標籤區域權重值設定原則。

(1) 平行式空間規劃：由於標籤區域皆為獨立存在，彼此間不互相影響，因此於此規劃下之標籤區域之權重設定僅參照各標籤重要性即可。

(2) 巢狀式空間規劃：由於一個標籤區域中（稱為外層標籤區域），其所含括之內容除了字彙外，尚包含其它標籤區域（稱為內層標籤區域），導致部分文件內容可能會同時由一個外層與多個內層標籤區域所含括。本研究乃認為此些重疊區域中之網頁內容，主要是以最內層標籤區域之標籤特性呈現，因此完全賦予最內層之標籤區域；然而，由於此內層標籤區域為多層標籤區域所含括，故此內層標籤區域的意義將會被適當地增強，因此於此內層區域亦需附加外層區域之權重值，以加強表現此標籤區塊字詞之重要性，是故此區域之權重計算方式如公式(1)所示，相關變數定義如下。

$W_{j,TR}$　分類依據標籤 j 位於網頁空間位置 TR 所分配之權重值，當中 $j \in \{T, B_1, B_2, ...\}$（請參照表 1）；$TR \in \{TypeIA, TypeIB, TypeIIA, TypeIIB, TypeIIIA, ...\}$（請參照圖 3 至圖 5 之空間規劃）。

α_i　含括於標籤 j（網頁空間位置 TR）之第 i 個外層標籤權重附加值。

$$W_{j,TR} = \sum_{all\ i}(1+\alpha_i)\cdot W_{i,TRC} \quad \text{where} \quad 0 \le \alpha_i \le 1 \tag{1}$$

(3) 混合式空間規劃：此空間規劃係同時運用平行式與巢狀式標籤區域處理原則，因此於權重計算方式與巢狀式空間規劃相同。

綜上所述，為避免型態名稱相同但所代表重要性不同之標籤區域，而視為是相等之情形（即將喪失許多有助於網頁文件分類之重要資訊）。因此本研究乃先將文字型標籤予以擷取，並配合標籤區域之空間規劃，以區分得此些在重要性上可能不同的標籤區域，並賦予對應之權重，其標籤區域權重設定彙整於表 2，當中標籤 B0 則代表無強調文字之標籤，但含括文字型態之標籤（如<font>等標籤）。是故，不同於以往之研究，本研究乃期望藉由考量網頁設計師於網頁標籤之文字撰寫與編排，並結合關鍵字擷取技術，以判定目標網頁之所屬類別。

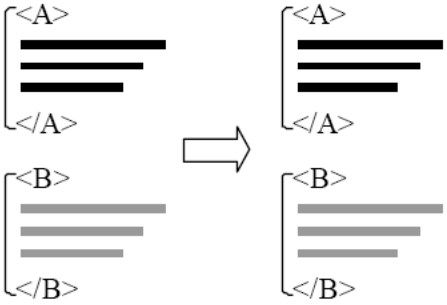

圖 3、平行式空間規劃 (Type I)

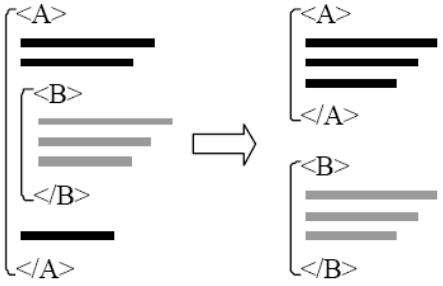

圖 4、巢狀式空間規劃 (Type II)

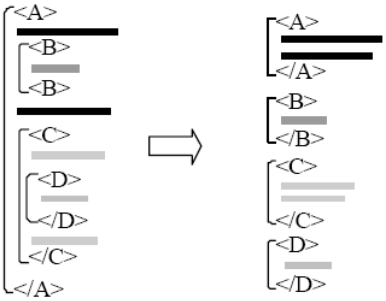

圖 5、混合式空間規劃 (Type III)

表 2、標籤區域權重設定彙整表

標籤代碼 (TR)	平行式空間規劃 (Type I)		巢狀式空間規劃 (Type II)		混合式空間規劃 (Type III)			
	A	B	A	B	A	B	C	D
T	$W_{Ti,TypeIA}$，當中<Title>僅出現於 Type I 之 A 區域，故僅考慮此區塊之權重							
B0	$W_{B0,TypeIA}$	$W_{B0,TypeIB}$	$W_{B0,TypeIIA}$	$W_{B0,TypeIIB}$	$W_{B0,TypeIIIA}$	$W_{B0,TypeIIIB}$	$W_{B0,TypeIIIC}$	$W_{B0,TypeIIID}$
B1	$W_{B1,TypeIA}$	$W_{B1,TypeIB}$	$W_{B1,TypeIIA}$	$W_{B1,TypeIIB}$	$W_{B1,TypeIIIA}$	$W_{B1,TypeIIIB}$	$W_{B1,TypeIIIC}$	$W_{B1,TypeIIID}$
B2	$W_{B2,TypeIA}$	$W_{B2,TypeIB}$	$W_{B2,TypeIIA}$	$W_{B2,TypeIIB}$	$W_{B2,TypeIIIA}$	$W_{B2,TypeIIIB}$	$W_{B2,TypeIIIC}$	$W_{B2,TypeIIID}$
B3	$W_{B3,TypeIA}$	$W_{B2,TypeIA}$	$W_{B3,TypeIIA}$	$W_{B3,TypeIIB}$	$W_{B3,TypeIIIA}$	$W_{B3,TypeIIIB}$	$W_{B3,TypeIIIC}$	$W_{B3,TypeIIID}$
B4	$W_{B4,TypeIA}$	$W_{B4,TypeIB}$	$W_{B4,TypeIIA}$	$W_{B4,TypeIIB}$	$W_{B4,TypeIIIA}$	$W_{B4,TypeIIIB}$	$W_{B4,TypeIIIC}$	$W_{B4,TypeIIID}$
AH	$W_{AH,TypeIA}$	$W_{AH,TypeIB}$	$W_{AH,TypeIIA}$	$W_{AH,TypeIIB}$	$W_{AH,TypeIIIA}$	$W_{AH,TypeIIIB}$	$W_{AH,TypeIIIC}$	$W_{AH,TypeIIID}$

3.2 網頁文件類別判定模組

本研究所提出之「網頁文件類別判定」模組乃以網頁標籤區域為分析基礎，先行蒐集當中文字型資料並擷取當中關鍵字詞，之後配合標籤區域權重分配模組針對各空間規劃下之標籤區域權重值，並配合領域關鍵字與類別之隸屬關係（需由領域專家先行建置），即可初步得知此目標網頁文件之所屬類別。首先將網頁文件類別判定模組所使用之符號定義如下：

符號定義

D_i	既有訓練網頁文件庫中之第 i 份網頁文件
$F(N_j,KW_i)$	標籤 j 所含括標籤區域中發生關鍵字 KW_i 之次數
G_k	第 k 種領域類別
KW_i	經關鍵字整併後，關鍵字集合之第 i 個關鍵字
$N(D_i,KW_i)$	訓練網頁文件 D_i 中發生關鍵字 KW_i 之次數
N_j	標籤 j 所含括之標籤區域，當中 $j \in \{T, B_0, B_1, B_2, ...\}$
$R(G_k,KW_i)$	關鍵字 KW_i 與領域類別 G_k 之隸屬係數
$Rlt_k'[WD_T]$	目標網頁文件 WD_T 與類別 G_k 之關係係數
$Rlt_k[WD_T]$	目標網頁文件 WD_T 與類別 G_k 之類別隸屬係數
WD_i	第 i 份網頁文件
WD_T	所考量之目標網頁文件

網頁文件類別判定模式可分數個步驟進行之。首先，於建構模式前需由系統管理者或領域專家建構一龐大關鍵字資料庫，此任務可運用孫銘聰、侯建良（2003）擷取訓練網頁文件之字串，歸納出現頻率較高且屬於關鍵字詞者為關鍵字，進而得到所有網頁文件之關鍵字集；這些關鍵字於不同網頁文件之發生頻率可整理如表 3。

表 3、各訓練網頁文件之領域關鍵字整併頻率摘要表

	KW_1	KW_2	……	KW_i	……
D_1	$N(D_1,KW_1)$	$N(D_1,KW_2)$	……	$N(D_1,KW_i)$	……
D_2	$N(D_2,KW_1)$	$N(D_2,KW_2)$	……	$N(D_i,KW_i)$	……
……	……	……	……	……	……
D_j	$N(D_j,KW_1)$	$N(D_j,KW_2)$	……	$N(D_i,KW_i)$	……

其次，利用領域網頁文件（如網頁新聞文件等資料）與類別之相關特性，將已知內容與類別之文件認定為訓練文件（即表 3 之 D_j），並根據侯建良、林峰興與畢威寧（2003）之方法論，將類別與關鍵字隸屬係數加以計算及精確化後，即可針對所有類別分別得到關鍵字 KW_i 與類別 G_k 之隸屬係數 $R(G_k,KW_i)$，其結果整理如表 4。

表 4、關鍵字與類別之類別隸屬係數表

	KW_1	KW_2	……	KW_i	……
G_1	$R(G_1,KW_1)$	$R(G_1,KW_2)$	……	$R(G_1,KW_i)$	……
……	……	……	……	……	……
G_j	$R(G_j,KW_1)$	$R(G_j,KW_2)$	……	$R(G_i,KW_i)$	……

本研究之網頁類別判定模式亦以此兩關係表為基礎，進而分為幾大步驟進行之；其流程包括網頁標籤區域之界定、網頁類別隸屬係數運算及類別隸屬係數之正規化等步驟。

步驟(A1)—界定網頁標籤區域

如同標籤區域權重分配模組所界定，本研究所將考慮之網頁文件資料先行以各網頁標籤進行區分，因此可將目標網頁文件 WD_j 乃劃分多個網頁標籤區域（$N_T, N_{B0}, N_{B1}, N_{B2}, N_{B3}, N_{B4}$）合併而成，如公式(2)所示。

$$WD_j = \{N_T, N_{B_0}, N_{B_1}, ..., N_{B_4}\} \quad (2)$$

其中，左式 WD_j 乃第 j 份網頁文件，而右式 N_T 至 N_{B4} 則為所考量網頁頭部標籤<title>所含括之網頁文字資料，以及網頁主體標籤<body>中文字標籤 B0 或文字強調標籤 B1 至 B4 所含括之網頁文字資料。

步驟(A2)—統計目標網頁文件關鍵字發生頻率

待界定網頁標籤區域後，利用此運用孫銘聰、侯建良（2002）關鍵字擷取方法，擷取目標網頁文件 WD_T 中各網頁標籤區域所包含之關鍵字，其結果整理如表 5。

表 5、目標網頁文件 WD_T 之關鍵字發生頻率表

WD_T	KW_1	KW_2	……	KW_i	……
N_T	$F(N_T,KW_1)$	$F(N_T,KW_2)$	……	$F(N_T,KW_i)$	……
N_{B0}	$F(N_{B0},KW_1)$	$F(N_{B0},KW_2)$	……	$F(N_{B0},KW_i)$	……
……	……	……	……	……	……
N_j	$F(N_j,KW_1)$	$F(N_j,KW_2)$	……	$F(N_j,KW_i)$	……

步驟(A3)—計算目標網頁文件與類別之關係

此步驟乃利用目標網頁文件關鍵字出現頻率（參照表 5）、領域關鍵字與類別關係之訓練資料庫（參照表 4），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（參照表 2），進而計算目標網頁文件 WD_T 與各類別 G_k 之關係係數 $Rlt_k'[WD_T]$，以初步判定此目標網頁文件類別偏向，並編列目標網頁文件 WD_T 與類別 G_k 之關係係數表；其分別如公式(3)及表 6 所示：

$$Rlt_k'[WD_T] = \frac{\sum_{all\,i}\sum_{all\,j}\sum_{all\,TR} R(G_k, KW_i)\cdot F(ND_j, KW_i)\cdot W_{j,TR}}{\sum_{all\,i}\sum_{all\,j}\sum_{all\,TR} N(KW_i, ND_j)\cdot W_{j,TR}}$$

$$where\ j \in \{T, B_0, B_1, B_2, ...\} \quad (3)$$

$$and\ TR \in \{TypeIA,\ TypeIB,\ TypeIIA,\ TypeIIB,\ TypeIIIA, ...\}$$

表 6、目標網頁文件與類別之關係係數表

	WD_1	WD_2	……	WD_T	……
G_1	$Rlt_1'[WD_1]$	$Rlt_1'[WD_2]$	……	$Rlt_1'[WD_T]$	……
……	……	……	……	……	……
G_k	$Rlt_k'[WD_1]$	$Rlt_k'[WD_2]$	……	$Rlt_k'[WD_T]$	……

此關係係數 $Rlt_k'[WD_T]$ 乃以關鍵字 KW_i 出現於所有網頁標籤區域 N_j 中之次數，並配合網頁空間規劃之權重值設定作為評斷比例，以及考量關鍵字與類別之隸屬係數 $R(G_k,KW_i)$，以初步獲得此關係係數。

步驟(A4)—計算目標網頁文件與類別之類別隸屬係數

由於目標網頁文件之關係係數總和不為 1（即 $\sum_{all\ k} Rlt_k'[WD_T] \neq 1$），因此，於此步驟乃將目標網頁文件 WD_T 與類別 G_k 之關係係數 $Rlt_k'[WD_T]$ 予以正規化，如公式(4)所示，可得另一係數（即目標網頁文件 WD_T 與類別 G_k 之類別隸屬係數，$Rlt_k[WD_T]$），其關係整理如表 7。此值愈大即代表目標網頁文件愈偏向該對應類別；若該值為 0，則表示目標網頁文件 WD_T 與類別 G_k 無相關。因目標網

頁文件 WD_T 之類別隸屬係數 $Rlt_k[WD_T]$ 為正規化後之係數，故其總和為 1（即 $\sum_{all\ k} Rlt_k[WD_T]=1$）。

$$Rlt_k[WD_T]=\frac{Rlt_k'[WD_T]}{\sum_{all\ k} Rlt_k'[WD_T]} \quad (4)$$

表 7、目標網頁文件與類別之類別隸屬係數表

	WD_1	WD_2	……	WD_T	……
G_1	$Rlt_1[WD_1]$	$Rlt_1[WD_2]$	……	$Rlt_1[WD_T]$	……
……	……	……	……	……	……
G_k	$Rlt_k[WD_1]$	$Rlt_k[WD_2]$	……	$Rlt_k[WD_T]$	……

經過上述四個步驟之推論，即可獲知目標網頁文件之隸屬類別。亦即根據此網頁文件之標籤區域與網頁空間規劃結果，並利用關鍵擷取技術，本研究即可判定目標網頁文件之類別。

3.3 鏈結網頁關聯程度推導模組

由於單一網頁文件無法涵蓋所有欲陳述之知識，故網頁設計者常利用目標網頁中之網頁鏈結標籤（即<a href>）以建立超連結，並使使用者由網路上之某一個節點跳到另一個網路上的節點（即可以由一份 HTML 文件跳到另一份 HTML 文件），以獲取更多相關知識，因此目標網頁與所鏈結之網頁尚具備高度關聯性之關係。

有鑑於此，本研究乃以目標網頁中網頁鏈結標籤，並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（參照表 2），擷取當中各網頁鏈結標籤之權重值，以及對應鏈結網頁之標題文字（即擷取<title>標籤之內文），進而評定各鏈結網頁與目標網頁關聯程度高低，建立不同關聯等級之區隔，以選定關鍵等級之鏈結網頁，並修訂目標網頁之類別。首先說明鏈結網頁關聯程度推導模組所使用之符號，並說明此模組之推導過程。

符號定義

β_j	相關程度遞減排序後之第 j 個鏈結網頁的修正權重係數
LWD_t	目標網頁文件之第 t 個鏈結網頁
$M[LWD_t]$	目標網頁文件 WD_T 與鏈結網頁 LWD_t 之關聯係數
$MRlt_k[WD_T]$	目標網頁文件 WD_T 與領域類別 G_k 之修正類別隸屬係數（即利用鏈結網頁關聯程度修正類別隸屬係數 $Rlt_k[WD_T]$）
$N(WD_T[TITLE])$	目標網頁文件 WD_T 之<title>標籤區域所包含之字數
$Rlt_k[WD_j']$	為選擇之前 S 個鏈結網頁中，排序為第 j 個鏈結網頁 WD_j' 與類別 G_k 之類別隸屬係數
WD_j'	依關聯係數 $M[LWD_t]$ 遞減排序後之第 j 個鏈結網頁

步驟(B1)—計算各鏈結網頁之關聯係數

首先，定義目標網頁與各鏈結網頁之關聯係數為其與對應鏈結網頁中標題文字字元重複比例，並結合鏈結網頁之鏈結標籤區域所分配之權重計算而得。因此，記算目標網頁 WD_T 與其第 t 個鏈結網頁 LWD_t 之標題文字重複字數比例，並配合標籤區域之權重值計算其與鏈結網頁 LWD_t 之關聯係數 $M[LWD_t]$，如公式(5)所示。

$$M[LWD_t]=\frac{N(WD_T[TITLE]\cap LWD_t[TITLE])}{N(WD_T[TITLE])}\cdot W_{AH,TR} \quad (5)$$

當中，$N(WD_T[TITLE]\cap LWD_t[TITLE])$ 係代表目標網頁與第 t 個鏈結網頁之標題文字重複字數比例。

步驟(B2)—制定各鏈結網頁預選等級

依據關聯係數 $M[LWD_t]$ 之計算結果予以排序，依此排序結果可制定預選等級。如選定排序前 S

個鏈結網頁為修訂目標網頁類別之層級，即應賦予各鏈結網頁對應之權重值 β_j（j=1,…,S），以修訂目標網頁之類別。

步驟(B3)—修正目標網頁與類別之類別隸屬係數

選定前 S 個鏈結網頁後，由關聯係數最高之鏈結網頁指定權重為 β_1、其次 β_2，依此類推直至 β_s（排序為前之權重值大於等於後者）。本研究乃依此權重修正目標網頁與各領域類別之類別隸屬係數，以求得修正後類別隸屬係數 $MRlt_k[WD_T]$，如公式(6)所示。此外，定義公式(6)中之 $Rlt_k[WD_j^{'}]$ 為選擇之前 S 個鏈結網頁中，排序為第 j 個鏈結網頁 $WD_j^{'}$ 與類別 G_k 之類別隸屬係數。

$$MRlt_k[WD_T]=\frac{Rlt_k[WD_T]+\sum_{j=1}^{S}\beta_j\cdot(Rlt_k[WD_j^{'}])}{1+\sum_{j=1}^{s}\beta_j}\quad where\ 0\le\beta_s\le\beta_{s-1}\le\ldots\le\beta_1\le 1 \tag{6}$$

此公式可用以修訂表 7，以提升計算目標網頁與類別之隸屬係數正確性；因此，修正後目標網頁之類別隸屬係數表如表 8 所示。完成上述各模組推論後，即可判定未分類網頁文件之隸屬類別。

表 8、修正後目標網頁文件與類別隸屬係數表

	WD_1	WD_2	……	WD_T	……
G_1	$MRlt_1[WD_1]$	$MRlt_1[WD_2]$	……	$MRlt_1[WD_T]$	……
……	……	……	……	……	……
G_k	$MRlt_k[WD_1]$	$MRlt_k[WD_2]$	……	$MRlt_k[WD_T]$	……

4. 網頁文件分類系統

根據第三章所提出之網頁文件分類模式，本研究乃發展一套以網頁標籤區域為基礎之網頁文件分類系統。於以標籤區域為基之網頁文件分類系統中，本研究乃分別就「網頁文件使用者」及「系統管理者」進行各功能模組之介紹，以說明此系統於實際應用之運作方式，如圖 6 所示。

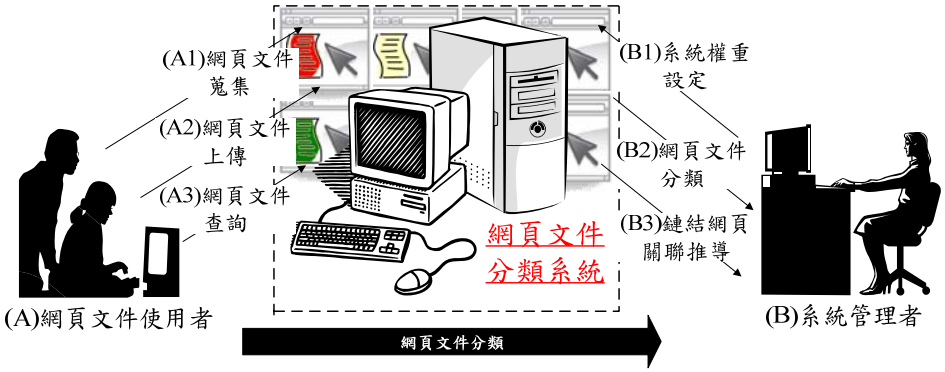

圖 6、網頁文件分類系統之應用流程

為使本系統能有效達成網頁文件分類，系統管理者須先設定系統權重，待設定完畢，系統即可開始運作。接著，網頁文件使用者需蒐集欲分類之網頁文件，並利用系統將蒐集之網頁文件上傳，上傳後，系統管理者利用系統之網頁文件分類功能，將欲分類之網頁文件勾選，送出後，系統即計算網頁文件之類別隸屬係數及網頁文件分類，分類結果分為二種，若隸屬係數大於門檻值，則係數愈大愈趨近該類別，若隸屬係數小於門檻值，則系統建議網頁文件使用者進行鏈結網頁關聯程度推導，以獲取更精確之網頁文件分類結果。最後，網頁文件使用者可透過網頁文件查詢功能檢視所上傳之網頁文件與分類結果，完成本系統之應用。

■ **系統管理者設定系統權重值**

首先，在利用系統分類網頁文件前，系統管理者須先針對網頁文件分類設定權重值、門檻值以及關鍵字與類別之類別隸屬係數（如圖 7、圖 8 所示），因本系統係以網頁標籤區域為分析基礎，

因此設定網頁標籤區域權重分配便有其重要性，於是本研究將所考慮之網頁文件資料先行以各網頁標籤進行區分，使系統管理者能立即對類別判定與關聯程式推導結果作出適當反應（亦即直接於系統中進行標籤區域權重、網頁類別隸屬係數與鏈結網頁關聯係數之參數修改），進而使網頁文件分類更能符合系統管理者之需求。

圖 7、「標籤區域權重分配」修改介面

圖 8、「網頁文件類別判定門檻值」修改介面

■ **網頁文件使用者蒐集網頁文件資料**

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」為驗證案例，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 9 所示，以下載「疑似飛碟在南華大學上空出沒」之網路新聞為案例，下載完畢後，即可利用本系統進行網頁文件分類。

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 10 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為「疑似飛碟在南華大學上空出沒.htm」、類別為「資訊」、關鍵字為「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為「疑似飛碟在南華大學上空出沒.htm」之檔案（詳見圖 10 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，以作為網頁文件分類之分析基礎。

圖 9、新聞「疑似飛碟在南華大學上空出沒」

圖 10、「網頁文件維護模組」上傳功能

■ **系統管理者執行網頁文件類別判定**

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，即可獲知目標網頁文件之隸屬類別（如圖 11、圖 12 所示）。

圖 11、網頁文件類別判定模組

圖 12、目標網頁文件類別判定（隸屬係數大於 0.5）

■ **系統管理者執行鏈結網頁關聯程度推導**

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」(系統預設為 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（如圖 13 所示）。

■ **網頁文件使用者查詢已分類之網頁文件**

待系統管理者將網頁文件分類完畢後，網頁文件分類資訊即儲存於資料庫中，網頁文件使用者可隨時透過「網頁文件查詢功能」查詢所上傳之網頁文件是否分類完成，「網頁文件查詢功能」乃提供網頁文件使用者查詢已上傳之網頁文件資料，以方便網頁文件使用者瞭解系統內各項網頁文件資料之維護結果。若查詢之網頁文件為「未判定」即網頁文件尚未分類，則網頁文件使用者仍可點選「詳細資料」查看所上傳之文件內容（如圖 14 所示）。

圖 13、鏈結網頁關聯程度推導

圖 14、「網頁文件查詢」網頁文件詳細資料

5.系統案例驗證與評估

為驗證本系統之網頁文件分類成果，本研究乃以「Yahoo 奇摩新聞」之網頁文件為探討案例。其中，驗證過程可分為訓練與測試二大階段，以下即針對系統驗證方式說明、評估指標定義與驗證結果分析等四步驟進行說明。

5.1 系統驗證方式說明

為驗證本研究所提出之方法論與系統績效，首先乃自十類新聞網頁文件類別，即「資訊類（G1）」、「社會類（G2）」、「財經類（G3）」、「體育類」(G4)、「健康類」(G5)、「教育類（G6）」、「藝文類（G7）」、「影劇類（G8）」、「旅遊類（G9）」與「生活類（G10）」等共計 1470 份網頁文件中，挑選出 350 份網頁文件作為訓練資料（各類別 35 份訓練資料）以建立關鍵字與類別之關係係數等資料。之後，由此十大類別中，各類隨機挑選兩份文件（共計 20 份）作為測試資料，利用訓練階段所修正之「關鍵字與類別之關係係數」，並推論此 20 份測試網頁文件之隸屬類別，並藉由觀察系統所推論之網頁文件類別是否符合該新聞網頁文件之實際類別，以確認本研究所提方法論之正確性。待完成上述之第一階段之系統績效驗證後，於第二階段中，本研究將剩餘之 1100 份新聞網頁文件分為 10 個週期持續匯入系統中，即每週期皆再匯入 110 份新聞網頁文件（各類別 11 份新聞網頁文件）；於各週期中乃利用前述 20 份測試新聞網頁文件重新進行網頁文件類別推論，以分析系

統於不同訓練網頁文件數量下之長期學期趨勢。

5.2 評估指標定義

為有效比較採用不同判定指標之推論結果，本研究將以「分類召回率」、「分類正確率」與「類別隸屬係數」此三項量化指標進行系統績效驗證，藉以檢視本研究推論目標網頁文件類別與其實際網頁文件類別之相符程度。分類召回率乃為一相對比例值，為「實際類別與推論類別相符之類別個數」與「實際類別個數」之比例（即 $R_i = \frac{m_i}{n_i}$）；另外，分類正確率則為一相對比例值，為「實際類別與推論類別相符之類別個數」與「推論類別個數」之比例（即 $A_i = \frac{m_i}{c_i}$）；最後，類別隸屬係數乃代表「系統推論目標網頁文件於實際類別隸屬係數」；此外，類別隸屬係數之平均值乃為一相對比例值，其代表「系統推論目標網頁文件於實際類別隸屬係數之總和」與「系統推論網頁文件總數」之比例（即 $AveWDG = \frac{\sum_{i=1}^{NT} WDG_i}{NT}$），期望藉由此項指標評估系統推論之目標網頁文件於各類別偏好與實際類別偏好間平均符合程度值。

5.3 系統驗證結果分析

本研究乃將系統驗證結果分為「第一階段驗證結果分析」與「第二階段驗證結果分析」兩大項目。於「第一階段驗證結果分析」項目中，乃以新聞網頁測試系統進行「網頁文件類別判定」與「網頁文件鏈結程度推導」之正確性，以瞭解系統擷取新聞網頁之內容、系統初期網頁文件類別結果之績效散佈狀況。而於「第二階段驗證結果分析」項目中，本研究乃於各測試週期新增訓練網頁文件（即訓練用具分類代表性之新聞網頁文件），以評估不同訓練資料量下系統進行「網頁文件類別判定」與「網頁文件鏈結程度推導」功能之績效。以下即以「第一階段驗證結果分析」、「第二階段驗證結果分析」與「驗證結果整體分析」等三主題說明本研究之驗證結果。

A 第一階段驗證結果分析

(A-1) 網頁文件類別判定（20 份指標網頁文件）

在 350 份新聞網頁文件作為訓練資料之基礎下，系統針對 20 份指標網頁文件之網頁文件類別判定平均召回率與為 95%（標準差為 0.2236），而網頁文件類別判定平均準確率為 95%（標準差為 0.2236），而網頁文件類別判定平均類別隸屬係數為 60%（標準差為 0.1585）；其中，指標網頁文件之推論結果與指標網頁文件之理想結果完全符合之網頁文件共 19 份，佔總測試網頁文件之 95%。而驗證結果之網頁文件類別判定召回率、準確率與類別隸屬係數的分佈趨勢如圖 15、16、17 所示。

(A-2) 網頁文件鏈結程度推導（20 份指標網頁文件）

在 350 份新聞網頁文件作為訓練資料之基礎下，系統針對 20 份指標網頁文件之鏈結網頁關聯程度推導平均召回率與為 95%(標準差為 0.2236)，而鏈結網頁關聯程度推導平均準確率為 95%(標準差為 0.2236)，而鏈結網頁關聯程度推導平均類別隸屬係數為 63%（標準差為 0.1681）；其中，指標網頁文件之推論結果與指標網頁文件之理想結果完全符合之網頁文件共 19 份，佔總測試網頁文件之 95%。而驗證結果之鏈結網頁關聯程度推導召回率、準確率與類別隸屬係數的分佈趨勢如圖 18、19、20 所示。

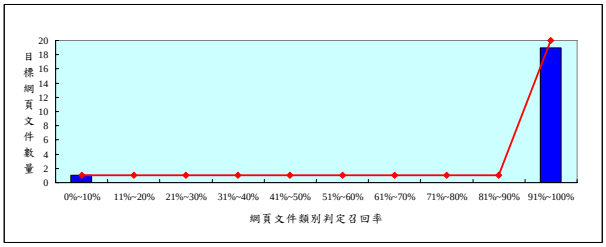

圖 15、第一階段網頁文件類別判定召回率

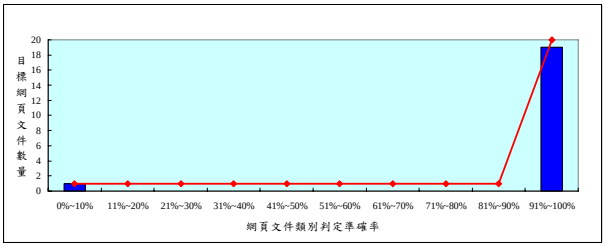

圖 16、第一階段網頁文件類別判定準確率

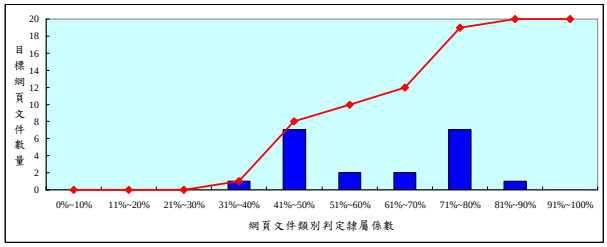

圖 17、第一階段網頁文件類別判定隸屬係數

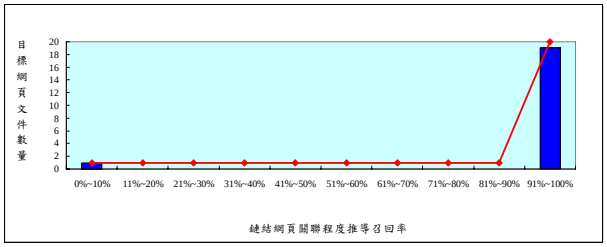

圖 18、第一階段鏈結網頁關聯程度召回率

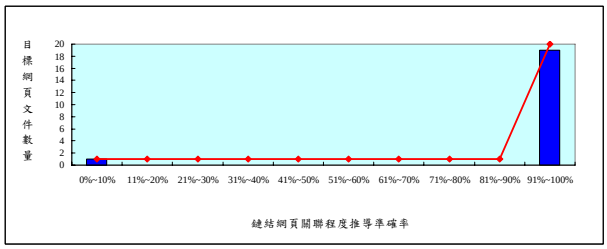

圖 19、第一階段鏈結網頁關聯程度準確率

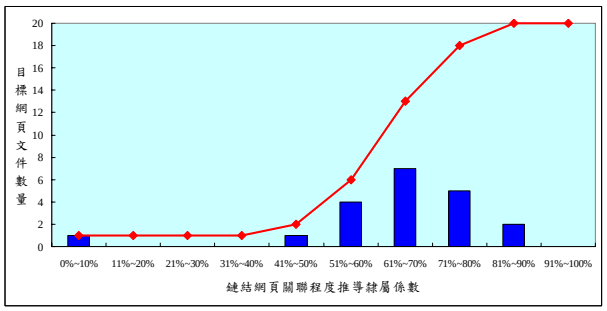

圖 20、第一階段鏈結網頁關聯程度隸屬係數

(B) 第二階段驗證結果分析

(B-1) 第二階段網頁文件類別判定之指標網頁文件結果

由圖 21 可知，以每週期增加 110 份訓練用網頁文件為單位，平均每週期網頁文件類別判定召回率與準確率之整體平均成長率皆為 0.50%；而判斷網頁文件類別判定類別隸屬係數之整體平均成長率為 1.70%，故可知網頁文件類別判定模組分類能力良好，且網頁文件類別判定模組具學習能力。

(B-2) 第二階段鏈結網頁關聯程度推導之指標網頁文件結果

由圖 22 可知，以每週期增加 110 份訓練用網頁文件為單位，平均每週期鏈結網頁關聯程度推導召回率與準確率之整體平均成長率皆為 0.5%；而判斷鏈結網頁關聯程度推導類別隸屬係數之整體平均成長率為 1.60%，故可知鏈結網頁關聯程度推導模組分類能力良好，並呈現穩定的推論能力。

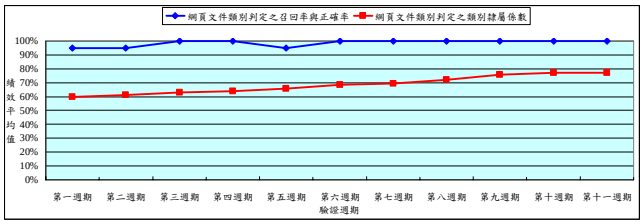

圖 21、各驗證週期網頁文件類別判定之績效分佈趨勢

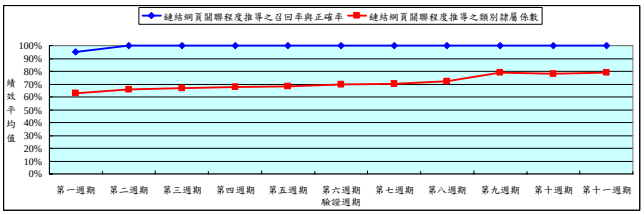

圖 22、各驗證週期鏈結網頁關聯程度推導之績效分佈趨勢

5.4 驗證結果整體分析

綜合二階段之驗證分析結果，可觀之無論評估網頁文件類別結果之績效散佈狀況，或評估不同訓練資料量下系統進行「網頁文件類別判定」與「網頁文件鏈結程度推導」功能之績效狀況，皆可看出本系統之有效性。於「第一階段驗證結果分析」中即可看出，本系統於初期網頁文件資料量為 350 份時，其「正確率」與「召回率」即可達到 95%，而類別隸屬係數之平均也可準確至 63%。而於「第二階段驗證結果分析」中，則可看出隨著不同週期數，本系統也隨著訓練網頁文件之增加而成長，於網頁文件類別判定時，其「正確率」與「召回率」可提升至 98.64%，且類別隸屬係數之平均也提升至 68.55%。此外，於鏈結網頁關聯程度推導時，其「正確率」與「召回率」已達至 99.55%，而類別隸屬係數之平均也升至 70.91%。此部份證明了本研究所建置之系統可行性，且只要不斷訓練更多正確之網頁文件資料，本系統便會更加成長。

綜合兩階段之驗證成效後，可將各項驗證指標之相關結果整理如表 9。由表 9 可知，各項驗證指標之收斂前每週期平均成長率及整體每週期平均成長率皆為正數，且各項驗證指標皆於十個週期內呈現收斂狀態，因此，以本研究所選驗證個案（即 Yahoo 新聞網頁文件）為例，當本系統使用

約 790 份至 900 份訓練用網頁文件時，可讓系統之各項推論績效提升至 95%以上的水準，故本研究所建置之系統可有效應用於網頁文件分類判定，並準確地依據判定結果給予使用者分類建議。

表 9、各項驗證指標成長率之彙整表

驗證指標	整體平均值	收斂週期	收斂前每週期平均成長率	整體每週期平均成長率
網頁文件類別判定召回率	98.64%	第五週期	1.25%	0.50%
網頁文件類別判定正確率	98.64%	第五週期	1.25%	0.50%
網頁文件類別判定類別隸屬係數	68.55%	第九週期	2.00%	1.70%
鏈結網頁關聯程度推導召回率	99.55%	第三週期	2.50%	0.50%
鏈結網頁關聯程度推導正確率	99.55%	第三週期	2.50%	0.50%
鏈結網頁關聯程度推導類別隸屬係數	70.91%	第九週期	2.00%	1.60%

6. 結論

由於現有之網頁文件類別判定技術未能考量以下兩點問題：(1)各標籤所含括之文字意義不同（如主題文字之標籤區域可擷取關鍵字詞較少，然該標籤區域卻是代表該網頁文件最重要之資訊）；(2)相同種類之標籤但位於不同位置標籤區域，其所包含文字區塊之重要性亦不盡相同之網頁文件設計方式，皆視各標籤區域之內容為同等重要資訊。因此，本研究乃建構一套「以標籤區域為基之網頁文件分類模式」，期望藉由考量標籤屬性與標籤位置，以建構標籤區域權重分配機制，並配合關鍵字擷取技術及網頁鏈結之關聯特性，以判定目標網頁文件之隸屬類別，進而達成網頁文件分類之目的。

參、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，「啟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為基礎之網頁分類器」，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.

出席國際學術會議心得報告

100 年　　02 月　　01 日

計畫編號	NSC 99-2221-E-343 -004
計畫名稱	以標籤區域為基之網頁文件分類模式
出國人員姓名 服務機關及職稱	楊士霆 南華大學資訊管理學系 助理教授
會議時間地點	2011/1/23~2011/1/25, Indonesia (Bali)
會議名稱	The 2011 International Conference on Asia Pacific Business Innovation & Technology Management (APBITM 2011)
發表論文題目	An Additional Component for Webpage Classification Technology

一、參加會議經過與心得

此次 APBITM 2011 研討會乃安排於印尼（Indonesia）之峇里島（Bali）舉辦，配合研討會主辦單位之行程規劃與可行機位安排，個人與國內學者（包含**銘傳大學林進財教授**、**政治大學楊亨利教授**、**東華大學陳啟斌教授**、**義守大學紀宗利教授**、**高雄應用科技大學張嘉倩教授**、**淡江大學阮聘如教授**等人）於 1/22 上午 6 點，即出發前往桃園國際機場，進行登記作業且於上午 09:35 由桃園國際機場起飛，並於當地時間 1/22 下午 14:35 抵達印尼峇里島之伍拉-賴國際機場（Ngurah Rai International Airport），台北與峇里島並無時差；此程搭機時間總計 5 小時，故辦理入境手續後搭車前往 Grand Mirage Resort Hotel 飯店（亦為 APBITM 2011 研討會之會議場所），稍做休息後，即於飯店周遭海域與附近商家逛逛，瞭解當地風俗民情與觀賞美麗的海岸。

次日（1/23），由於研討會晚間才得以報到，因此利用本日之上半天至峇里島中部參觀，以瞭解峇里島傳說等主題，並參觀展覽館，以獲知峇里島歷史旅程，之後前往烏布傳統市場，瞭解當地風俗民情與飲食習慣，最後於晚上五點則前往 Grand Mirage Resort Hotel 飯店，完成報到手續（如圖 1 及圖 2 所示）。

此次研討會總計發表篇數約為 220 篇左右，議程數為 22 個左右。大會正式行程日期為 1/23~1/25 三日，正式發表日期乃為 1/24~1/25 兩日。個人乃於 1/23 下午五點已至會場，並完成報到手續（如圖 1 及圖 2 所示）。本次研討會內容乃安排與此次會議主題相關之企業、人資、科技及創新等專題演講與論文發表，再依不同論文主題每天分至 2 至 4 個時段，以及 3 至 4 個左右平行 Session 進行發表。個人的論文被安排於發表日第一天（1/24）的下午場次（編號 Session M2-3）「Knowledge & Technology Management」發表，於發表後其他學者亦表示對此研究的高度興趣，詢問本研究之網頁分類技術與其他研究之差異，個人並作完整回答，互動甚佳。此外個人亦參加多場與研究興趣較相關之發表場次，並對於其他學者發表內容提出詢

問，對於知識管理、作業研究等課題觸發新的研究靈感（如圖 3 至圖 6 所示）。

除會議發表時間外，在其他交流活動時，個人與國際/國內學者（如銘傳大學林進財教授、淡江大學阮聘如教授等人）亦有良好交流，於此次研討會所認識之多數先進乃屬國內「**台灣作業研究學會**」及「**中華決策科學學會**」（如圖 7 及圖 8 所示），因此可瞭解許多國際/國內工業工程、資訊管理學者之研究方向，並規劃未來合作之可能作法，收穫極大，此對於個人學術經歷尚屬資淺而言，乃一大助益。

圖 1、抵達 APBITM 2011 會場並註冊(1)


圖 2、抵達 APBITM 2011 會場並註冊(2)


圖 3、論文報告與研討(1)

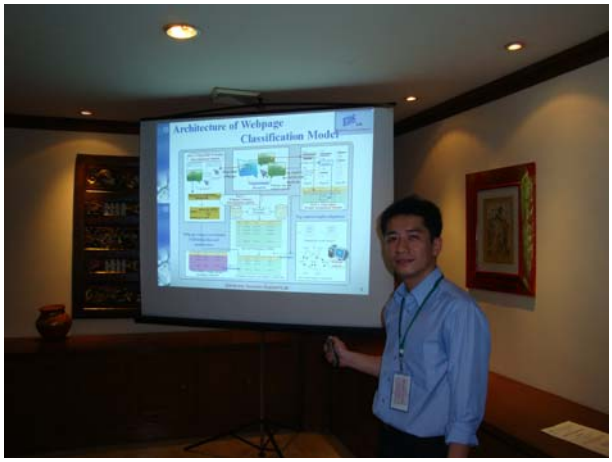

圖 4、論文報告與研討(2)

圖 5、論文報告與研討(3)

圖 6、論文報告與研討(4)

圖 7、與國內學者合影(1)

圖 8、與國內學者合影(2)

待研討會圓滿結束後（1/25），個人與國內學者一行人則利用 26 日上半天前往特產店參觀，以瞭解當地之出產品（如咖啡、胡椒粉、咖哩、雞湯醬、蝦餅等）。之後乃搭車前往機場並搭機飛返回桃園國際機場，結束此次 APBITM 2011 學術研討活動。

三、建議

此次會議中的各項活動安排都可發現主辦單位頗為用心，對於遠道造訪之學者給予多項貼心之服務，為國內學校爭取主辦國際型研討會可加以參考之長處。然而，雖然主辦單位之用心可見，但由於此次研討會乃於 Grand Mirage Resort Hotel 飯店舉辦，因此需受限於飯店之場地限制（如各議程場地較為狹小、領取資料流程、網路提供與連線、以及休息區之規劃皆不甚完美），此可提供國內學者於辦此類大型學術研討會之借鏡。

整體而言，本次大會舉辦頗為用心，個人於此行收穫豐富，且結識多位國際學者，希望能於未來建立更長遠的交流與合作。

四、攜回資料名稱及內容

1. 研討會論文集：含議程集 1 本、論文摘要集 1 本及研討會光碟 1 片。
2. 國內外學者學術交流名片。

An Additional Component for Webpage Classification Technology

S. T. Yang
Department of Information Management, Nanhua University, Taiwan
stingyang@mail.nhu.edu.tw

Abstract -Webpage classification technology with high accuracy can improve the efficiency for Internet users to search required knowledge and to save lots of knowledge-searching time. Differing from previous researches, this paper explores an additional component for webpage classification. That is, concerning complexity of webpage structure, this paper analyzes the webpage layout including tag attributes and tag-region locations designed in webpage to develop an algorithm for webpage classification. Therefore, based on webpage layout analysis, the text contained in specific tag-regions can be identified. Also, the keywords extracted from each tag-region are weighted according webpage layout analysis and then the categories of the target webpage can be determined. Furthermore, based on the hyperlink tag, the similar webpage with higher correlation can be collected to re-determine target webpage categories. In addition to the webpage classification algorithm, a web-based webpage classification system is developed to demonstrate feasibility of the proposed model. The attempt of this research is to propose an addition component and concept (i.e., webpage layout, tag attributes and tag-regions analysis) for webpage classification technology to improve the effectiveness of webpage classification.

Keywords: Tag-region, Webpage Classification, Webpage Design, Keyword Extraction, Knowledge Management

I. INTRODUCTION

With the advancement of Internet technologies, the number of Internet users is increasing and the amount of information online has growth explosively. As browsing information or files on the Internet has become one of important channels for knowledge acquisition, how to effectively manage Internet information/files to assist the users in efficiently absorbing and utilizing required information has become an important issue. On the basis of this issue, many technologies for webpage classification have been developed. Since webpage contents contain texts, pictures or films, most researches analyze and classify these kinds of data for categorization. Also, some researches maintain domain keywords in database as a basis for determination of webpage categories. Furthermore, as the tags appearing in pairs (i.e., tag-region) contain words of certain segmentation (e.g., <title></title > and <h1></h1>) in webpage, some researches apply the standardized programming pattern of UML (Unified Modeling Language) and HTML (Hypertext Markup Language) used by webpage creators/designers to analyze webpage tags for webpage classification. In case of insufficient data for webpage classification, other researches employ hyperlinks contained in webpage for such purpose (i.e., the relevant and additional information on hyperlink webpage are analyzed for webpage classification). The AS-IS model of webpage classification is as shown in Fig. 1.

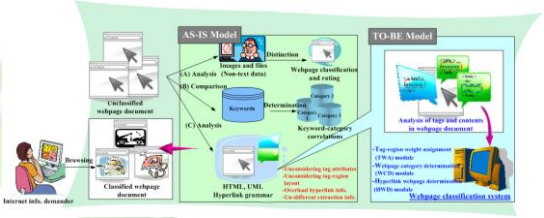

Fig. 1: The As-Is and To-Be models of webpage classification

Different from previous webpage classification methodologies, this paper concentrates on webpage design characteristics (including tag attributes and tag-region layout, etc.) for webpage classification. That is, this paper analyzes tag attributes (e.g., head tag <title> always contains more representative words with respect to the target webpage) and also considers tag-region layout (i.e., tags of the same type located in different tag-region layout contains words of different importance) to assign the corresponding weights for all considered tags. After that, this paper employs keyword extraction technology to extract all the keywords contained in different tag-regions and given the corresponding weights. Finally, to avoid problems of insufficient or excessive analysis information, this paper establishes a hyperlink webpage screening mechanism to collect the hyperlink webpage with higher correlations to slightly modify or adjust the categories of the target webpage. The TO-BE model of this paper is shown in Fig. 1.

II. LITERATURE REVIEW

Concerning the webpage classification issue, most researches focus on analyzing the texts, tags and hyperlinks contained in webpage.

(A) Webpage text information analysis

Previous researches extract keywords for webpage classification based on webpage text information [2,3,11]. Besides the above extraction of keywords, the semantic in webpage text is also analyzed [14]. The SRG (Semantic Relationship Graph) is constructed according to the spatial scale that could be searched under the guidance of combinative table and association list, and then the Naive Bayesian classifier is used to develop a semantic relationship graph based multi-relationship Naive

Bayesian classifier [8]. Such classifier removes unnecessary characteristics and relationships according to the analytical results of semantic relationship graph to avoid generating uncorrelated associations [1,5].

(B) Webpage tag information analysis

Lim et al. [4] proposes UML and HTML grammars or tag characteristic in webpage documents as the analytic data of webpage classification. The extracted data are taken as analysis characteristics and data of webpage classification for further studies [6]. In addition, based on the DOM (Document Object Model) tag-tree structure, a webpage can be segmented into small tag-regions. Each tag-region can be displayed in the browser by visualized types corresponding to a specific nested combination of tag-pairs. The profitability of tag-regions for webpage classification is varying among visual types caused by the web authoring convention [12].

(C) Webpage hyperlink information analysis

Furnkranz [9] classifies webpage documents considering the hyperlink ensembles in webpage. The classification data can be obtained from the text of the target webpage and the hyperlink webpage and used to classify the target webpage effectively. The OEM (Object Exchange Model) is employed to identify webpage categories [15]. In this methodology, the number of hyperlink of the target webpage is calculated and the contents of hyperlink webpage are converted into Node Similarity, Edge Similarity and Structural Similarity to obtain the similarity degree to classify the similar webpage into the same category.

III. A MODEL FOR TAG-REGION ANALYSIS AND WEBPAGE CLASSIFICATION

The webpage classification model proposed in this paper is based on analysis of tag attributes and tag-regions to search for text contained in tag-regions and extract the corresponding keywords. Based on tag attributes and specific tag-region layout of webpage, the corresponding weight values are assigned to various tag-regions. Therefore, according to keywords extracted from various tag-regions and weights assigned for tag-regions, the categories of target webpage can be determined. Finally, hyperlink tag (i.e., <a href>) is used to search for hyperlink webpage with higher correlation to modify the categories of the target webpage. Therefore, this model can be divided into three kernel modules including "tag-region weight assignment (TWA) module" (as shown in Part 1 of Fig. 2), "webpage category determination (WCD) module" (as shown in Part 2 of Fig. 2) and "hyperlink webpage determination (HWD) module" (as shown in Part 3 of Fig. 2).

Part 1: Tag-region weight assignment (TWA) module

In TWA module, this paper acquires the tags (i.e., tag extraction mechanism) correlated with webpage classification (i.e., which contain text data); then, the weight assignment of tag-region in different tag-region spatial layout (i.e., tag-region location analysis mechanism) are discussed.

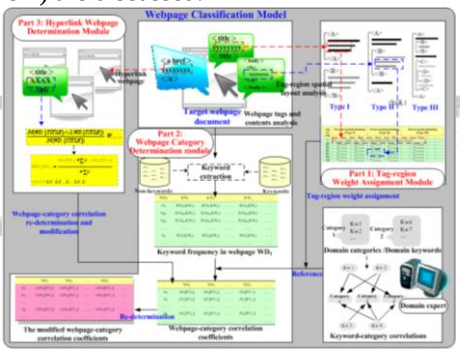

Fig. 2: Architecture of webpage classification model

(A) Tag extraction mechanism

In HTML, all tags have their respective purposes, and contents contained in tag-regions often reflect the attributes of those tags. For example, head tags (including <head>, <title>, <bgsound>, <meta>, <style>, <script>, etc.) are mainly used to define segments of such setting values as format, form, name, Script language and pattern list of the target webpage. Among which, head tag <title> is containing the subject of the webpage, and designers often employ tags <h1> to <h6> to display and highlight subject of different sizes. Webpage body tag (i.e., <body>) contains all texts, pictures and other multi-media files to be displayed. Concerning display of webpage texts, webpage designers often employ bold/italic tags (including <b>, , <cite>, and etc.), tabulating tags (including <ol>, <li>, <ul>, <dl>, <dt>, <dd>, etc.) and quoted text tag (i.e., <blockquote>) to emphasize importance of terms displayed in webpage. In a webpage, texts contained in different tag-regions denote different significance and importance. Therefore, webpage classification model in this paper takes text data in webpage as analysis basis. The webpage head tag and webpage body tag are mainly utilized, and text tone strengthening tags are also considered to serve as the basis for webpage classification. Webpage tags to be employed are summarized in Table 1.

Table 1: List of tags as the classification basis

Tag types	Tag Names	Text highlight tags	
Head	<title>	Subject tags (T)	<h1>, <h2>, <h3>, <h4>, <h5>, <h6>…
Body	<body>	Bold tags (B_1)	<b>, …
		Italic tags (B_2)	<address>, <cite>, <dfn>, , <i>….
		Tabular tags (B_3)	, , , <dl>, <dt>, <dd>…
		Quotation tags (B_4)	<blockquote>…
Hyperlink	<a href>	Hyperlink tag (AH)	

(B) Tag-region location analysis mechanism

Following tag extraction as classification basis, the extracted tags of the same type but located in different positions may contain texts of different significance. In order to differentiate these tags, this module analyzes the spatial layout of tag-region [7], and assigns weights to tag-regions located in different spatial layout.

As HTML has the function of spatial planning of webpage, this module further discusses the relationship between tag-regions and webpage spaces. That spatial planning of tag-regions can be divided into three types including (1) Parallel spatial planning, (2) Nested spatial planning and (3) Mixed spatial planning (as shown in Fig. 3 to Fig. 6) and the principle of tag-region weight

assignment for spatial planning are described as follows.

(1) Parallel spatial planning: As tag-regions are all independent from each other, weight assignment of tag-regions is only referred to tag attributes.

(2) Nested spatial planning: As one tag-region (i.e., external tag-region) contains not only contents but also other tag-regions (i.e., internal tag-region), some contents of the webpage may be contained in one external tag-region and several internal tag-regions simultaneously. The innermost tag-region should also be provided with weight values of external tag-regions to enhance text significance therein. Weight calculation of the inner tag-region is shown in Equation (1) and symbols used in this mechanism are defined as follows.

$W_{j,TR}$ The weight value of the j'th tag located in TR, $j \in \{T, B_1, B_2, \ldots\}$ (see Table 1) and $TR \in \{TypeIA, TypeIB, TypeIIA, TypeIIB, TypeIIIA, \ldots\}$ (see spatial planning in Fig. 3 to Fig. 6).

α_i The added weight value of the i'th external tag contained in the j'th tag (located in TR).

$$W^*_{j,TR} = \left[1 + \sum_{all\ i}(\alpha_i)\right] \cdot W_{j,TR} \quad \text{where} \quad 0 \le \alpha_i \le 1 \qquad (1)$$

(3) Mixed spatial planning: This spatial planning applies the principles of the above two spatial planning; so that, weight assignment is the same as that of nested spatial planning.

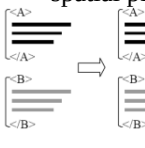

Fig. 3: Parallel spatial planning (Type I)

Fig. 4: Nested spatial planning (Type II)

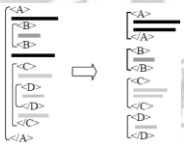

Fig. 5: Mixed spatial planning (Type III)

To avoid tag-regions with the same form and name but different significance as the same ones (in such case, critical information for webpage classification may be lost), this module extracts tags first, along with spatial planning of tag-regions to differentiate those tag-regions which may have different importance, and then assign the corresponding weight values. The weight values of tag-regions can be summarized in Table 2.

Table 2: List of weight values of tag-regions

Tag No. (TR)	Parallel spatial planning (Type I)		Nested spatial planning (Type II)		Mixed spatial planning (Type III)			
	A	B	A	B	A	B	C	D
T	$W_{T,TypeIA}$; where head tag <title> is only existed in the A location of Type I							
B0	$W_{B0,TypeIA}$	$W_{B0,TypeIB}$	$W_{B0,TypeIIA}$	$W_{B0,TypeIIB}$	$W_{B0,TypeIIIA}$	$W_{B0,TypeIIIB}$	$W_{B0,TypeIIIC}$	$W_{B0,TypeIIID}$
B1	$W_{B1,TypeIA}$	$W_{B1,TypeIB}$	$W_{B1,TypeIIA}$	$W_{B1,TypeIIB}$	$W_{B1,TypeIIIA}$	$W_{B1,TypeIIIB}$	$W_{B1,TypeIIIC}$	$W_{B1,TypeIIID}$
B2	$W_{B2,TypeIA}$	$W_{B2,TypeIB}$	$W_{B2,TypeIIA}$	$W_{B2,TypeIIB}$	$W_{B2,TypeIIIA}$	$W_{B2,TypeIIIB}$	$W_{B2,TypeIIIC}$	$W_{B2,TypeIIID}$
B3	$W_{B3,TypeIA}$	$W_{B3,TypeIA}$	$W_{B3,TypeIIA}$	$W_{B3,TypeIIB}$	$W_{B3,TypeIIIA}$	$W_{B3,TypeIIIB}$	$W_{B3,TypeIIIC}$	$W_{B3,TypeIIID}$
B4	$W_{B4,TypeIA}$	$W_{B4,TypeIB}$	$W_{B4,TypeIIA}$	$W_{B4,TypeIIB}$	$W_{B4,TypeIIIA}$	$W_{B4,TypeIIIB}$	$W_{B4,TypeIIIC}$	$W_{B4,TypeIIID}$
AH	$W_{AH,TypeIA}$	$W_{AH,TypeIB}$	$W_{AH,TypeIIA}$	$W_{AH,TypeIIB}$	$W_{AH,TypeIIIA}$	$W_{AH,TypeIIIB}$	$W_{AH,TypeIIIC}$	$W_{AH,TypeIIID}$

Part 2: Webpage category determination (WCD) module

Based on weight values of tag-regions in different spatial planning obtained TWA module, as well as the correlations between keywords and categories established by domain expert in advance, the categories of the target webpage can be determined in WCD module. Before classifying the target webpage, the keywords should be established by domain experts in advance. Then, the training webpage D_i (i.e., the webpage of known contents and categories) are used to calculate the frequencies $N(D_i,KW_i)$ of keywords existing in each training webpage (as shown in Table 3). After that, by using correlations between training webpage (such as news webpage) and domain categories, the correlation coefficient $R(G_i,KW_i)$ between keyword KW_i and category G_i can be established in Table 4 [10].

Table 3: Frequency of keyword in training webpage

	KW_1	KW_2	.....	KW_i	.....
D_1	$N(D_1,KW_1)$	$N(D_1,KW_2)$	.....	$N(D_1,KW_i)$	.....
D_2	$N(D_2,KW_1)$	$N(D_2,KW_2)$	.....	$N(D_2,KW_i)$	.....
.....	.....	.....	.....	.....	.....
D_j	$N(D_j,KW_1)$	$N(D_j,KW_2)$	.....	$N(D_j,KW_i)$	.....
.....	.....	.....	.....	.....	.....

Table 4: The keyword/category correlation coefficients

	KW_1	KW_2	.....	KW_i	.....
G_1	$R(G_1,KW_1)$	$R(G_1,KW_2)$	.....	$R(G_1,KW_i)$	.....
G_2	$R(G_2,KW_1)$	$R(G_2,KW_2)$	.....	$R(G_2,KW_i)$	.....
.....	.....	.....	.....	.....	.....
G_j	$R(G_j,KW_1)$	$R(G_j,KW_2)$	.....	$R(G_j,KW_i)$	.....
.....	.....	.....	.....	.....	.....

The WCD module in this paper is also based on these two established Tables. The details for webpage classification are introduced as follows.

Step (C1): Definition of webpage tag-regions

As discussed in TWA module, all the webpage considered in this paper can be segmented through tags. So that, the each webpage WD_j can be divided into several tag-regions (N_T,N_{B0},N_{B1},N_{B2},N_{B3},N_{B4}) (see Equation (2)).

$$WD_j = \{N_T, N_{B_0}, N_{B_1}, \ldots, N_{B_4}\} \qquad (2)$$

Where N_T~N_{B4} represent webpage texts contained in webpage head tag <title>, as well as webpage texts contained in text tag B_0 or text highlight tags B_1 to B_4 in body tag <body>.

Step (C2): Calculation of frequencies of keywords in the target webpage

After webpage tag-regions are defined, keywords extraction technology [13] is adopted to extract keywords contained in tag-regions in the target webpage WD_T.

Step (C3): Calculation of relationship coefficient between target webpage and categories

Based on the derived keyword frequencies in the target webpage, keyword-category correlation coefficients (see Table 4) and tag-region weight assignment (see Table 2), the relationship coefficient $Rlt_k'[WD_T]$ between the target webpage WD_T and category G_k can be obtained via Equation (3) to preliminarily determine the categories of the target webpage.

$$Rlt_k'[WD_T] = \frac{\sum_{all\ i}\sum_{all\ j}\sum_{all\ TR} R(G_k, KW_i) \cdot F(D_j, KW_i) \cdot W_{j,TR}}{\sum_{all\ i}\sum_{all\ j}\sum_{all\ TR} F(D_j, KW_i) \cdot W_{j,TR}} \qquad (3)$$

where $j \in \{T, B_0, B_1, B_2, \ldots\}$

and $TR \in \{TypeIA, TypeIB, TypeIIA, TypeIIB, TypeIIIA, \ldots\}$

Step (C4): Calculation of correlation coefficients between target webpage and categories

As the sum of relationship coefficients of target webpage

is not equal to 1. In Step (C4), the relationship coefficient $Rlt_k'[WD_T]$ between target webpage WD_T and category G_k should be standardized (as shown in Equation (4)) to obtain the webpage-category correlation coefficient $Rlt_k[WD_T]$. If the webpage-category coefficient is greater, the target webpage approaches the corresponding category. On the other hand, if the value is equal to zero, the target webpage WD_T have no relation to category G_k.

Part 3: Hyperlink webpage determination (HWD) module

As single webpage cannot cover all knowledge to be described, webpage designers often use hyperlink tags (i.e., <a href>) in target webpage to build webpage hyperlink. Based on the hyperlink, Internet users can link from the target webpage to another webpage for acquisition of more relevant knowledge. Therefore, the relationship existed between target webpage and linked webpage should be discussed.

This module uses the hyperlink tags in target webpage to derive the hyperlink webpage with higher correlation for redetermination and modification of the categories of the target webpage. Firstly, referring to the weight assignment of tag-regions (see Table 2), the weight values of all hyperlink tags and subject words of the corresponding hyperlink webpage can be acquired. Secondly, the relationship between each hyperlink webpage and the target webpage can be calculated and ranked to select the hyperlink webpage within predefined selection degree. After that, the categories of target webpage can be re-determined accordingly.

Step (D1): Calculation of correlation value of hyperlink webpage with respect to the target webpage

Firstly, the weight values of all hyperlink tags in the target webpage WD_T can be obtained from TWA module. Secondly, the repetition proportion of subject words between each hyperlink webpage LWD_t and WD_T can also be obtained. After that, the correlation value of LWD_t for WD_T can be derived via Equation (5).

$$M[LWD_t] = \frac{N(WD_T[TITLE] \cap LWD_t[TITLE])}{N(WD_T[TITLE])} \cdot W_{AH,TR} \quad (5)$$

Step (D2): Setting of selection degree of similar hyperlink webpage

All the hyperlink webpage in WD_T are ranked according to their correlation values $M[LWD_t]$ (in descent order). Also, the selection degree S should be defined in advance. The top S ranking hyperlink webpage (WD_j', where $j \le S$) are selected and the modification weight values β_j (j=1,…,S) of these selected webpage are assigned to re-determine the categories of the target webpage.

Step (D3): Re-determination of correlation coefficients of target webpage and categories

After top S ranking hyperlink webpage are selected, the hyperlink webpage with higher correlation values are given corresponding weight values (i.e., β_1, β_2,…, β_s, the weight value of the preceding one is greater than or equal to that of the following one). Based on these modification weight values, this module re-determines the correlation coefficients of target webpage and categories to obtain the modified correlation coefficient $MRlt_k[WD_T]$ (as shown in Equation (6)).

$$MRlt_k[WD_T] = \frac{Rlt_k[WD_T] + \sum_{j=1}^{s} \beta_j \cdot (Rlt_k[WD_j'])}{1 + \sum_{j=1}^{s} \beta_j} \quad (6)$$

$$where\ 0 \le \beta_s \le \beta_{s-1} \le \ldots \le \beta_1 \le 1$$

IV. WEBPAGE CLASSIFICATION SYSTEM

A web-based portal, namely webpage classification system, is developed for webpage classification over Internet. Under this system, the webpage documents could be maintained and the user authorities are properly managed so that the webpage classification results can be accurately provided to this developed staff.

Based on the user login information, the webpage classification system recognizes the user category (e.g., system administrator and common user) and provides the corresponding functions to the user. Under the system, the system administrator establishes the domain keywords with respect to the specified categories to database via keyword maintenance module as the foundation of system training (Fig. 6). Also, the system administrator can upload the webpage documents with given categories to the database via training webpage document upload function (Fig. 7). After that, the keyword-category correlations and webpage-category correlations can be established in system database. After uploading these training webpage documents, the webpage-category correlation coefficients of unclassified webpage documents uploaded by common users can be determined by system administrator through webpage classification function (Fig. 8 and Fig. 9). Furthermore, if all the correlations of the target webpage are not greater than predefined threshold, the system automatically recommends administrator to re-determine the webpage-category correlations via hyperlink webpage analysis function (Fig. 10). In addition, the system administrator can set weight values of tags according tag attributes, category determination threshold and hyperlink webpage selection threshold, etc. through the system parameter maintenance module (Fig. 11). Finally, the system administrator can maintain users' profiles and control users' authorities via user profile maintenance module. Under the platform, common user also can upload the webpage documents via webpage upload function, so that the webpage documents can be efficiently managed and shared. Also, common user can review or download all kinds of webpage documents provided by system administrator or other common users in the database via webpage search function (Fig. 12). After the webpage categories are determined, common users also can inquire webpage-category correlations of

target webpage via webpage search function (Fig. 13).

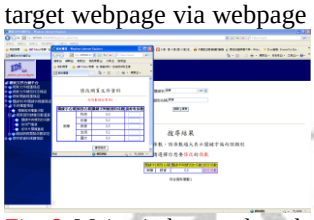

Fig. 6: Maintain keywords and keyword-category correlations

Fig. 7: Upload training webpage documents to the system

Fig. 8: Select webpage documents for classification

Fig. 9: Results of webpage classification

Fig. 10: Re-determine webpage-category correlations

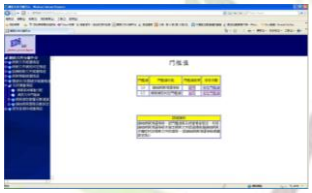

Fig. 11: Set hyperlink webpage selection threshold

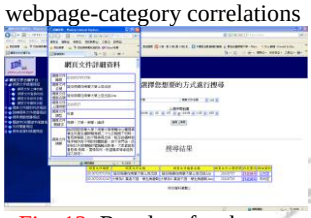

Fig. 12: Results of webpage inquiry (1)

Fig. 13: Results of webpage inquiry (2)

V. CONCLUSION

Different form technologies for webpage classification, this paper analyzes tag attributes and tag-region layout in webpage to develop an algorithm for webpage classification including tag-region weight assignment (TWA) module, webpage category determination (WCD) module and hyperlink webpage determination (HWD) module. In TWA module, tag attributes and tag-region layout designed in webpage are analyzed to assign weight values to the corresponding tag-regions. In WCD module, the keyword extraction technology is employed to extract keyword contained in each tag-region and the corresponding weights are given to determine the webpage-category correlations. In HWD module, the hyperlink webpage with higher correlations with respect to the target webpage are used to re-determine and modify the webpage categories. The attempt of this research is to improve the accuracy and efficiency of webpage classification by concerning the characteristics of webpage design. Also, the proposed webpage classification algorithm can assist the information demanders to efficiently and effectively search the required information over the Internet; so that, lots of researching energy and time can be reduced.

REFERENCES

1. A. Fujino, N. Ueda, and K. Saito, "A hybrid generative/discriminative approach to text classification with additional information," *Information Processing and Management*, vol. 43, pp. 379-392, 2007.
2. C. C. Yang, J. Yen, and H. Chen, "Intelligent internet searching agent based on hybrid simulated annealing," *ELSEVIER Journal on Decision Support System*, pp. 269-277, 2000.
3. C. M. Chen, H. M. Lee, and Y. J. Chang, "Two novel feature selection approaches for web page classification," *Expert Systems with Applications*, vol. 36, no. 1, pp. 206-272, 2009.
4. C. S. Lim, K. J. Lee, and G. C. Kim, "Multiple sets of features for automatic genre classification of web documents," *Information Processing & Management*, vol. 41, no. 5, pp. 1263-1276, 2005.
5. E. Youn, and M. K. Jeong, "Class dependent feature scaling method using naive Bayes classifier for text datamining," *Pattern Recognition Letters*, vol. 30, no. 5, pp. 477-485, 2009.
6. H. Artail, and F. Kassem, "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, 2008.
7. H. C. Hsu, "The web page classifier based on progressive tagged-region analysis," *Department of Information Engineering, Tamkang University, Master Thesis*, 2000.
8. H. Chen, H. Liu, J. Han, X. Yin, and J. He, "Exploring optimization of semantic relationship graph for multi-relational Bayesian classification," *Decision Support Systems*, vol. 48, no. 1, pp. 112-121, 2009.
9. J. Furnkranz, , "Hyperlink ensembles: A case study in hypertext classification," *Information Fusion*, vol. 3, no. 4, pp. 299-312, 2002.
10. J. L. Hou, and F. H. Lin, "A document and user matching model via document keyword analysis," *Journal of Computer Information Systems*, vol. 44, no. 4, pp. 1-15, 2004.
11. J. T. Horng, and C. C. Yeh, , "Applying genetic algorithms to query optimization in document retrieval," *Information Processing and Management*, vol. 36, pp. 737-759, 2000.
12. L. C. Sung, "Progressive analysis scheme for web document classification," *Department of Information Engineering, Tamkang University, PhD Dissertation*, 2006.
13. M. T. Sun, and J. L. Hou, "The architecture and models for security reasoning in an EDMS," *Journal of the Chinese Society of Industrial Engineers*, vol. 20, no. 4, pp. 305-316, 2003.
14. S. Tan, and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," *Expert Systems with Applications*, vol. 34, pp. 2622-2629, 2008.
15. Y. H. Kuo, and M. H. Wong, "Web document classification based on hyperlinks and document semantics," *PRICAI 2000 Workshop on Text and Web Mining*, pp. 44-51, 2000.

ACKNOWLEDGMENT

This research is partially supported by the National Science Council under project No. NSC 99-2221-E-343 -004.

國科會補助計畫衍生研發成果推廣資料表

日期:2011/08/09

國科會補助計畫	計畫名稱: 以標籤區域為基之網頁文件分類模式
	計畫主持人: 楊士霆
	計畫編號: 99-2221-E-343-004- 學門領域: 資訊系統

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：楊士霆	計畫編號：99-2221-E-343-004-
計畫名稱：以標籤區域為基之網頁文件分類模式	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	2	2	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	1	2	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	此計畫「以標籤區域為基之網頁文件分類模式」乃獲「2011 中華企業創新與經營學會全國論文競賽」-佳作獎。此外，大專生所參與之技術開發，亦獲 2011 年「工業工程與管理」學生專題論文「資訊系統組」決賽資格。

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估 ■達成目標 □未達成目標（請說明，以 100 字為限） □實驗失敗 □因故實驗中斷 □其他原因 說明：
2. 研究成果在學術期刊發表或申請專利等情形： 論文：■已發表 □未發表之文稿 □撰寫中 □無 專利：□已獲得 □申請中 ■無 技轉：□已技轉 □洽談中 ■無 其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以500字為限）

本研究之目標乃提昇網頁文件分類技術之正確率與效率性。對於資訊需求者而言，本研究則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。根據上述，本研究重點完成工作項目如下：

1.標籤區域權重分配方法論之建立：

(1)蒐集並回顧標籤區域解析、網頁空間規劃之相關研究與文獻

(2)以標籤區域與網頁空間規劃為基礎，建立標籤區域權重分配之方法論

2.建構網頁文件類別判定模式：

(1)蒐集並回顧網頁文件分類之相關研究與文獻

(2)建構以鏈結網頁為基礎之網頁文件類別修訂模式

3.系統功能模組建構：

(1)建構標籤區域權重分配模組

(2)建構鏈結網頁關聯程度推導模組

(3)建構網頁文件類別判定模組

4.案例驗證與成果分析

本研究以網路新聞文件為案例進行案例驗證。驗證過程所採用之新聞文件乃由「Google奇摩新聞」與「YAHOO奇摩新聞」等網路新聞社群所蒐集之網路文件為本研究之驗證資料，以確認本研究所提模式與技術之正確性、合理性。待模式驗證完成後，最後檢討實際成效與預期成果間之符合程度，並由分析評估瞭解本研究之未來發展與應用方向。