以 MMB 演算法改良中文網站自動分類系統的效能

Using MMB Algorithm to Refine the Performance of Chinese Web Site Automatically Classified System

駱思安　　李中彥　　徐俊傑
台灣科技大學資訊管理所　中國文化大學資訊管理所　台灣科技大學資訊管理所
alex.hero@msa.hinet.net　cylee@staff.pccu.edu.tw　cchsu@cs.ntust.edu.tw

摘要

每一個網站中包含著許許多多的文字, 分散在網站內的每一個網頁中, 而這些文字一部分是描述網站隸屬於何種類別, 另一部分則是與隸屬類別毫無關係的雜質。因此，如能有效地去除網站中的雜質文字, 即能成功地提昇中文網站自動分類的效能。本研究提出 WSACS(Web Site Automatically Classified System)，一個中文網站自動分類系統，有效地去除網站中的雜質文字，並採用 MMB(Multimembership Bayesian)的方式來推論網站的類別隸屬。

WSACS 有三大模組，知識建構模組採用應用程式滲透測試的方式, 探勘出網站中網頁的超鏈結架構和文句，並運用 CKIP 斷詞器將文句做適當的切割並賦予詞性, 僅留下詞性為名詞的詞彙讓去除贅詞和同義詞決定單元來過濾以產生網站詞集，最終運用計算 p_{ij} 和 $\overline{p}_{ij}$ 值的公式，用以產生網站詞集各自的 p_{ij} 和 $\overline{p}_{ij}$ 值；推論引擎模組以 MMB 為理論基礎來推論網站的類別隸屬；知識學習模組則在固定的時間內，自動學習詞彙、p_{ij} 和 $\overline{p}_{ij}$ 值，以確保推論知識的正確性。

關鍵詞：多重關係貝氏方法、知識管理、決策支援系統、網站分類、中文斷詞

Abstract

A Web site contains a lot of terms which are distributed in each Web page of the Web site. Some of these terms describe the characteristics of the Web site and can used to classify the Web site to a specific category. The others have no relationship to the Web site are ignored while performing the classification task. So, if we can eliminate the noisy terms, we can successfully improve the performance of Chinese Web site automatically classified system. In the research, a Chinese Web site hierarchical classification system is developed. In the system, the useful terms are extracted by using the p_{ij} and $\overline{p}_{ij}$ value. The knowledge base is then applied by the MMB approach to infer the individual probabilities of categories which the Web site belongs to. The category with the highest probability is selected by WSACS to the designated classification category.

In the system, there are three major modules. The first is the knowledge construction module. The module uses the Web mining to explore the Web page's hyperlink structure and sentences. Then, I try to cut the sentences with CKIP segmentation unit into different terms, and all non-noun terms are eliminated. The ambiguity terms are removed from noun terms and all synonyms are grouped. The result term set is used to calculate p_{ij} and $\overline{p}_{ij}$ values to construct the inference knowledge base. The second is the inference engine module. It uses the MMB approach along with the inference knowledge base to infer the Web site's classification probabilities. The third is knowledge learning module which provides a self-learning mechanism to update the inference knowledge base.

Keywords：Multimembership Bayesian approach、Knowledge Management 、 Decision Support System、Website Classify、Chinese Segmentation

一、前言

隨著網際網路的日新月異，網路上每天都充斥著許多的資料, 故瀏覽者必須要耗費相當多的時間和精力去做篩選和過濾的動作, 以便擷取出其真正需要的資料集。而知識管理的目標是「知識」，不是「資料」，所以在資料氾濫的時代中，需要的是組織過的資料及分類後的知識, 也因此現今探討分類技術議題(例如：文章分類、網頁分類和網站分類)的文章不斷地被各學界所提出。

而本研究將以網站分類作為研究的主軸，而網站內容含意分類的目的，是在對網站內容進行分門別類的「加值處理」，使性質相近的網站被放在相同或較類似的類別當中，如此網站群組不但易於被管理，而且可以不斷地被使用者快速瀏覽或萃取。傳統的網站分類工作需要利用大量人力，如此，不僅要耗去許多寶貴的時間，而且人工分類始終存在一個「標準如何界定」的嚴重問題，亦即不同的人會有不同的分類結果；根據研究，不論是人工建索引或是人工判定網站相似度等，其準確率大約只有「60%」。而利用電腦來做網站自動分類的實驗，從 1960 年代就已開始，並陸續有相關論文發表，自從 90 年代初期網際網路普及，自動分類更成了人人叫喊的「顯學」。而一套永續經營的網站自動分類系統，應具備有自動建議分類機置的功能，使系統在偵測出有越來越多無法分類的資訊時，能夠自動歸類近似特徵，建議產生新分類，這樣才能夠滿足當今使用者的需求。

而分類的動作是一門相當專業的技術和學問，且必須要搭配電腦的自動化處理，才能成為人性與科技結合的最佳典範，從全球相關領域的技術文獻中的經驗可得知，企業組織在規劃自動化文件內容分類的設計方法時，主要有三種導入法，分別是「樣本資料法」（Top Down）、「帶槍投靠法」（Bottom Up）和「科學式混合法」（Scientific Hybrid）[7]。

「樣本資料法」也就是傳統的人工分類法，先由參與的專業人員訂定分類架構，再讓他們依自己的專業認知，選出符合各個分類的文件數百篇，上傳系統之後，讓系統從這些文件中找出每個分類中的「重要詞彙」，而將這些詞彙聚集起來，即是某一個分類的「特徵集」，如此，日後若有新進文件，經過系統判讀，發現這篇新進文件本身的特徵結構與某分類的特徵集在比例上有一定的關係，即將之自動歸類於這一個分類當中。此法的優點為，結合專業人員的專業知識，使每個分類的特徵是由專業人員間接配合系統訂出的，也因此會注意到文件內容的上下文關係；面對大量資訊時，此法最符合工作所需，主觀上也較準確；而此法的缺點為，專業人員的專業素養不夠時，選出的樣本文件品質較差，故所有產出的分類特徵都將有問題；此法過程中有一個特徵詞彙除錯的動作，如果各專業人員沒有達成共識，各自有自己的認定方式時，每一個分類裡的特徵結構就會產生問題。

「帶槍投靠法」則是目前較炙手可熱的網站自動分類方式，不需要客戶的參與，而提供自動分類系統的廠商即主動提供一套完整的分類架構和每一個分類裡的特徵集。以「醫藥產業」為例，由於藥學已是一門百年學問，因此已有許多藥典的分類架構和專有詞庫可供參考，而且行之有年沒有爭議性，因此自動化系統是可以直接將之擷取引用的。此法的優點為，相較於前述的「樣本資料法」，這是較自動化的方法，將客戶端人力參與成本降至最低，因此能夠快速導入和使用；此法的缺點為，能不能配合客戶方面暨有的分類架構，即是考驗各家廠商的技術能力。沒有一套分類架構是能永遠適用於每一種產業的，因此這樣的方法，僅適合於一些專業並行之有年的產業，因為這些產業多數已經累積了非常多的詞典或專用的查詢分類等等，但其對於新式產業或以創新為主題產業則很難有廠商能配合。

「科學式混合法」則是綜合上述二法的特點所產生出的改良法，面對的情況是，假設一個企業客戶有非常多的特定資料內容，但既沒有可效法之分類架構、又沒有專業人員有專業知識可以主觀製定一套分類架構時，客戶與廠商要合作產生出一套新分類架構的過程，因此此法強調廠商需要有非常先進或客制化的技術能力，一般而言，需要的技術能力有三種：提供分類訓練工具、提供預設分類架構和提供自動建議分類機置。

二、相關研究和文獻探討

本研究所採用的 MMB 方式曾經於 1991 年在美國伊利諾理工學院用來發展 MEDAS 醫療診斷專家系統[14]，目的是希望在醫生診斷的過程中，由新病人的單一或多個病症，綜合先前其他病人的歷史病例，推論出新病人患病的機率，儘可能推論出更多的病症給醫生作為決策時的參考，以便協助醫生正確無誤地開藥方給新病人服用。而 WSACS 的目的和 MEDAS 相仿，故 MMB 的概念可用於 WSACS，即在同一時間內，運用所有網站類別群組個別的推論知識，自動計算出目標網站隸屬於第 H 層(最底層)的多目標類別之個別機率值，以便提供使用者作為網站分類時的參考，進而提昇網站正確分門別類的準確率。

現今一般的網站自動分類方式是把目標網站集先分成二個部分，其中一部分先做為訓練資料集(Training Data Set)，訓練資料的目的是用來將各個類別之特性探索出來，以便作為未來分類的依據所在，故此步驟需要在進行分類動作之前先進行；另一部分將做為測試資料(Testing Data)，而測試資料是用來測試分類依據的分類準確度，如圖 1 所示[24]。

網站自動分類的步驟，一開始要將訓練網頁進行文字處理(Text Processing)，而文字處理主要是將網頁中的文句資料轉換成為詞彙資料，而當網頁文句處理完中文斷詞和去除常用字(Stop-List)步驟之後，接下來要在文字處理步驟所產生的字詞集合中找出能代表網站的關鍵詞集，也就是找出能代表網站的特徵詞集，通常關鍵詞的萃取是利用詞彙在網站內容網頁中所出現的頻率(Term Frequency，簡稱 TF)來計算此詞彙在網站中的權重計算。

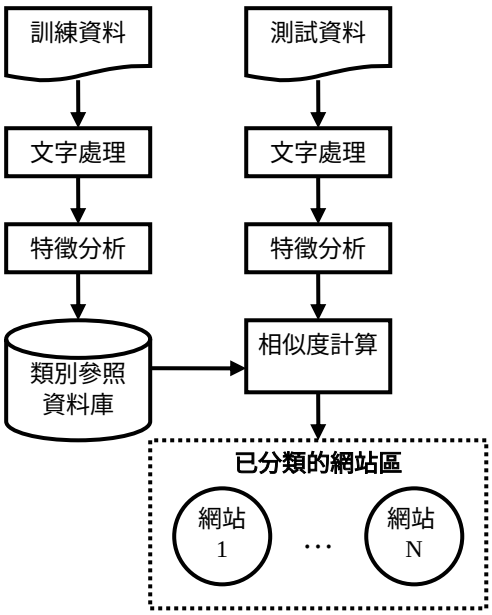

圖 1　常見網站自動分類的步驟

至於文件自動分類方面，早期文件被抽象化為關鍵字與重要性數字的關係後就可以套用到一般的機器學習與分類技術，自動文件分類器在近年來已有相當多學者投注其研究上，像是知名的貝氏機率模型[20][31][32]、支撐向量機(Support Vector Machine)[17][26][27][28][32]、及 KNN(K Nearest Neighbor) [23][30][31][32]。

「貝氏機率模型」是一個基於機率理論的分類方法。在特徵選取後，由已知文件計算出該特徵與該類別之間的條件機率關係，分類時藉由此機率關係計算文件屬於各類的機率，由其中選出機率最高的類別作為該文件的類別。但這樣的方法的表現並不理想，主要由於它對各特徵間的獨立假設。且容易因訓練文件中包含文件較多的大類而有所偏差，特徵對各類別的分別意義在這個方法中不容易顯現出來。

「SVM」它能夠將原有的訓練資料所在的空間 X 透過 Mercer 核心運算子(Kernel Operator)轉換成另一個更高維的空間。它的目標是自中找出一個最佳的分割超平面(Hyper-Plane)，這個超平面能夠達到將兩類點分得最開，也就是有最大的邊界(margin)。這個超平面僅是由訓練資料在空間中的點中與該平面距離 $1/|w|$ 的點決定，若僅以這些點訓練，會得到相同的支撐向量(Support Vector)。其可以獲得統計學習理論上依訓練資料所得的最佳結果[29]，但相較於其他方法它的計算量顯得相當龐大。「KNN」這個方法利用待分文件和其鄰近的文件相似，所以待分文件可以依鄰近文件的類別來判斷它的類別，其主要的缺點為計算可能相當費時[18]。

三、研究方法和系統架構

(一)、研究方法

WSACS 發展初期，將會廣泛參考國內外相關的論文，以及目前網際網路上實際獲取的資訊，也就是資料收集法，作為本研究的理論基礎所在；接著，採用實驗驗證法，依據本研究引用的 MMB 理論實作出一個 WSACS，並藉由實驗來驗證 WSACS 之可行性[1][2][3][4][12]。

(二)、WSACS 之系統架構

WSACS 包括有三大模組，分別是知識建構、推論引擎、知識學習模組，如圖 2 所示。而這三個模組具有執行上的順序限制，且彼此之間呈現一種相依相存的關係；亦即在知識建構模組建構好 MMB 推論知識庫後，推論引擎模組才能夠進一步運用此 MMB 推論知識，配合 MMB 的推論方式，推論出網站階層式類別決策資訊，而知識學習模組也要在推論引擎模組推論出最適當的網站階層式類別決策資訊後，才能著手進行詞彙、p_{ij}、$\overline{p_{ij}}$ 值的知識學習動作。知識建構模組當中的 MMB 推論知識庫，攸關著推論引擎模組推論的結果是否準確；而推論引擎模組所推論出的網站階層式類別決策資訊，則攸關著知識學習模組是否能有效地對詞彙、p_{ij}、$\overline{p_{ij}}$ 值進行知識學習的動作；而知識學習模組當中新生的詞彙、p_{ij}、$\overline{p_{ij}}$ 值品質是否良好，則牽動著知識建構模組當中的 MMB 推論知識品質是否變好。故 WSACS 中三個模組間的關係是環環相扣缺一不可的。

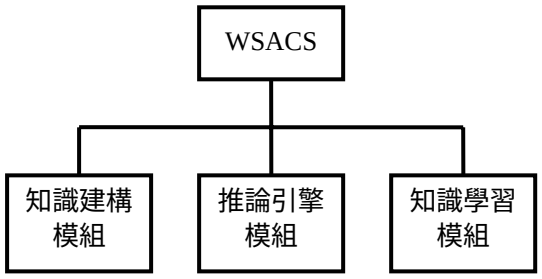

圖 2　WSACS 之系統概念圖

圖 3 為 WSACS 之運作流程圖，總共包括有三個模組、七個主要處理元以及三個延伸處理單元。其中，三個模組包括有知識建構模組、推論引擎模組以及知識學習模組；七個主要處理元，包括有知識來源處理元[6]、知識擷取處理元、知識過濾處理元[8]、知識表示處理元、MMB 推論引擎處理元、網站分類目錄辨識處理元和知識學習處理元；三個延伸處理單元，則包括有知識創新處理元、科技創新處理元以及知識價值處理元。

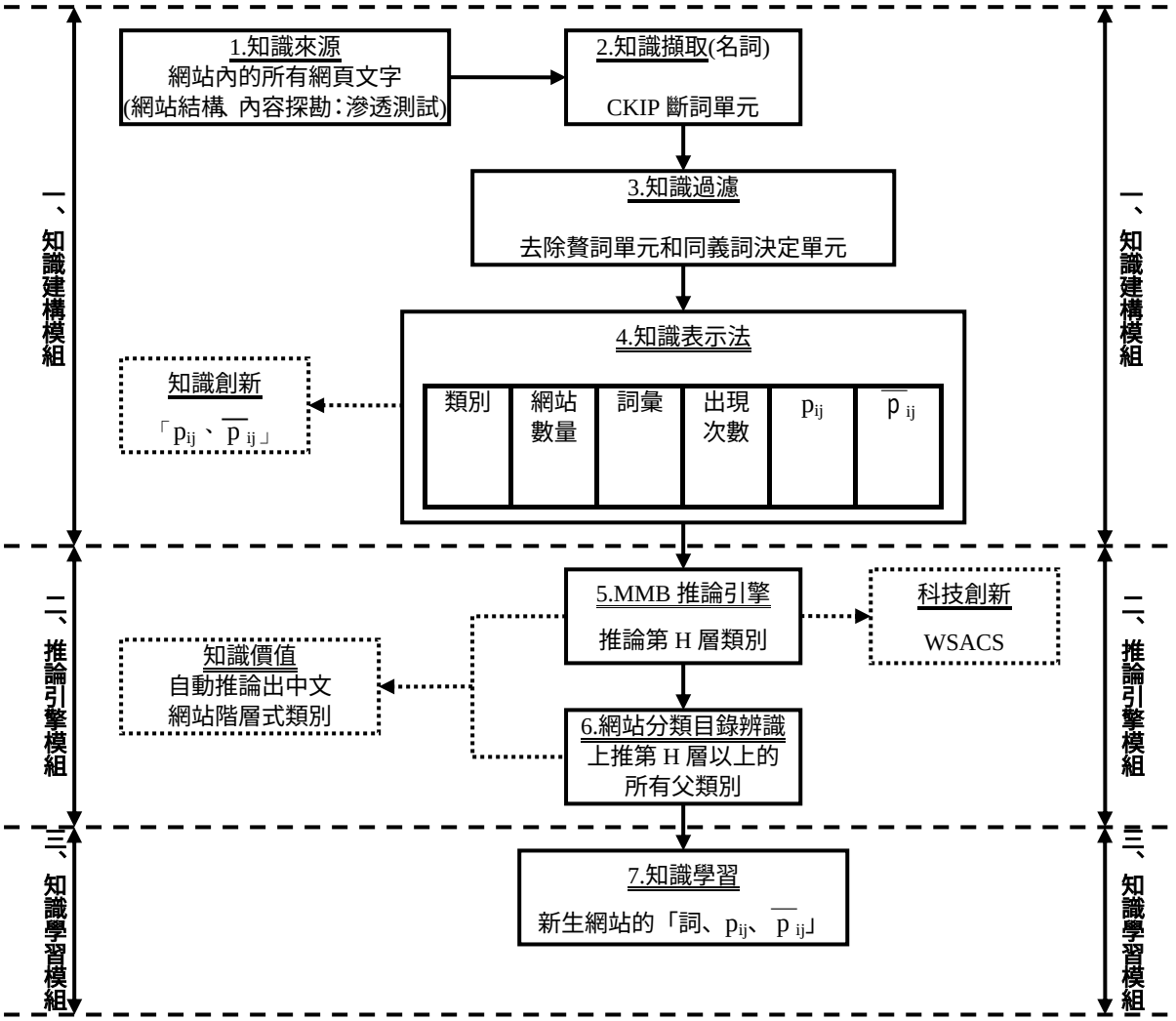

圖 3　WSACS 之系統架構圖

如圖 3 所示，WSACS 當中的每一個處理元之間都存在著牽一髮而動全身的關係，故網站分類人員只要缺少或省略其中任何一個處理元，都將有可能導致 WSACS 最終所推論出來的網站階層式類別決策資訊不夠準確。而 WSACS 當中較為重要的處理元，不外乎是知識來源、知識擷取、知識過濾以及知識學習四個處理元；其原因如下所示：由於倘若 WSACS 未能從一個品質佳的知識來源取得推論知識，儘管其採用再好的知識擷取與知識過濾技能都只是白費工夫，而 WSACS 由知識來源處理元取得知識後，若未經過知識擷取和知識過濾處理元的處理，則這些網站的資訊是無法成為其真正所需的 MMB 推論知識，而 WSACS 將有可能會因此推論出錯誤的決策資訊；此外，WSACS 假若使用了這些品質不佳的網站階層式類別決策資訊進行 MMB 推論知識學習的動作，則 MMB 推論知識庫的整體品質將會不斷地被破壞，終將面臨毫無參考價值的窘境。

WSACS 的第 H 層(最底層)MMB 推論知識庫是由兩種數值所建構而成，分別是 $P(W_j \mid B_i)$(以 p_{ij} 代表)和 $P(W_j \mid \overline{B}_i)$(以 $\overline{p}_{ij}$ 代表)。p_{ij} 代表在類別 B_i 裡的所有知識訓練網站樣本中詞彙 W_j 出現的機率值，而 WSACS 根據詞彙 W_j 在類別 B_i 裡的知識訓練網站樣本中所出現之次數，來標示「K_1」($K_1 \geqq 1$)，倘若詞彙 W_j 沒有出現，則標示「0」；而 $\overline{p}_{ij}$ 代表在類別 $\overline{B}_i$ 裡的所有知識訓練網站樣本中，詞彙 W_j 會出現的機率，WSACS 根據詞彙 W_j 在類別 $\overline{B}_i$ 裡的知識訓練網站樣本中所出現的次數來標示「K_2」($K_2 \geqq 1$)，倘若詞彙 W_j 並沒有出現，則標示「0」即可。而 MMB 推論知識庫當中，有關計算詞彙 p_{ij} 和 $\overline{p}_{ij}$ 值的公式之介紹及說明，則如公式(1)(計算 p_{ij} 值)和公式(2)(計算 $\overline{p}_{ij}$ 值)所示[14][16]。

$$p_{ij} = P(W_j \mid B_i) = \frac{\text{非 "0" 的數量}}{\text{屬於類別} B_i \text{的網站數}}, 1 \leq i \leq m, 1 \leq j \leq Q \quad (1)$$

$$\overline{p}_{ij} = P(W_j \mid \overline{B}_i) = \frac{\text{非 "0" 的數量}}{\text{不屬於類別} B_i \text{的網站數}}, 1 \leq i \leq m, 1 \leq j \leq Q \quad (2)$$

在介紹完有關計算詞彙 p_{ij} 和 $\overline{p}_{ij}$ 值的公式之後，接下來將進一步地介紹有關 MMB 公式的相關說明，如公式(3)所示。

$$P(B_i \mid W_1, W_2, ..., W_n) = \frac{P(B_i)\times P(W_1 \mid B_i)\times ..\times P(W_n \mid B_i)}{P(B_i)\times P(W_1 \mid B_i)\times ..\times P(W_n \mid B_i)+(1-P(B_i))\times P(W_1 \mid \overline{B}_i)\times ..P(W_n \mid \overline{B}_i)} \quad (3)$$

其中，WSACS 將 $P(W_j \mid B_i)$ 以 p_{ij} 代表 $P(W_j \mid \overline{B}_i)$ 以 $\overline{p_{ij}}$ 代表、$P(B_i \mid W_1, W_2, ..., W_n)$ 代表目標網站包含有詞彙 $W_1, W_2, ..., W_n$ 後屬於類別 B_i 的後天機率值，而 $P(B_i)$ 則是目標網站屬於類別 B_i 的先天機率值，WSACS 將 $P(B_i)$ 的值預設為 0.5，這代表一開始在沒有任何預設立場的情況下，WSACS 欲判別的目標網站屬於第 H 層各類別 ($B_1, B_2, ..., B_m$) 的個別機率值均相等。其中，1≦i≦m；1≦j≦n。

(三)、WSACS 之相關分類議題

WSACS 目前所提及的相關分類議題共有三個，分別為階層式類別、多重類別和新增類別。其中，「階層式類別」[5]為 WSACS 的加值功能，亦即當 WSACS 已判別出網站的最底層類別時，WSACS 便會藉由事先建構好的網站分類目錄架構來上推出目標網站完整的階層式類別隸屬；「多重類別」為當網站藉由 MMB 公式運算完畢後，有一個以上的候選類別項之後天機率值高於門檻值 0.8，則這些候選類別項(多筆)皆有可能為網站的類別隸屬；如圖 4 所示，「新增類別」為當沒有一個候選類別項之後天機率值高於 0.6，則 WSACS 會將其分入「其他」類別當中，待日後其他類別當中的網站數量增加到一定的程度，WSACS 未來將會嘗試藉由這些目標網站的關鍵詞集之 TF(Term Frequency)值和 IDF(Inverse Document Frequency)值[19][21][22]來尋找詞彙集中的相關性，進一步去克服新增類別的瓶頸。

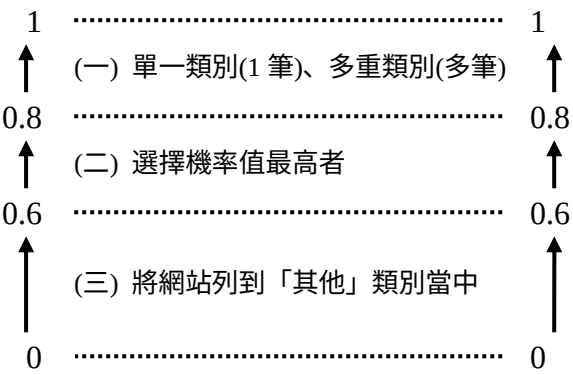

圖 4　WSACS 之系統概念圖

四、網站自動分類方式之綜合探討

(一)、WSACS 之系統運作流程介紹

本文的實證探討將會分成三個部分來切入說明，分別是「知識建構模組」、「推論引擎模組」和「知識學習模組」，而詳細的介紹和說明如下所示：

＊知識建構模組之範例

首先，網站分類人員必須先決定 WSACS 的網站分類目錄架構之階層數 H(H>0)為何[9]，然後開始著手建構第 H 層的網站分類目錄架構，如圖 5 所示。

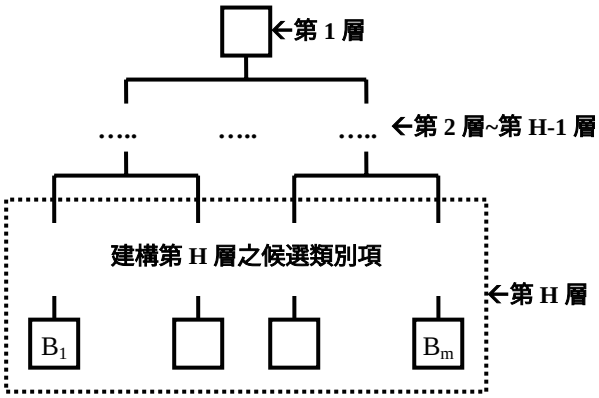

圖 5　WSACS 中 H=3 的網站分類目錄架構

其次，WSACS 將以類別 B_1(職棒)、類別 B_2(大學)和類別 B_3(醫院)這三個類別為例，其中，取類別 B_1(職棒)的知識訓練網站樣本數共 30 個、類別 B_2(大學)共 50 個和類別 B_3(醫院)共 70 個，如表 1 所示。

表 1　WSACS 中第 H 層(H=3)的各類別知識訓練網站樣本範例

B_i：類別	網站名稱
B_1：職棒	1 中信鯨,…,30 兄弟象
$\overline{B}_1$：{B_2, B_3}	1 三民書局, ... ,120 榮總醫院
B_2：大學	1 文化大學,…,50 台灣科技大學
$\overline{B}_2$：{B_1, B_3}	1 文化大學,…,100 榮總醫院
B_3：醫院	1 長庚醫院,…,70 榮總醫院
$\overline{B}_3$：{B_1, B_2}	1 中信鯨,…,80 台灣科技大學

最終，運用計算 p_{ij} 和 $\overline{p}_{ij}$ 的公式，開始建構第三層類別 B_1、類別 B_2 和類別 B_3 個別的 MMB 推論知識，表 2、表 3、表 4 為這三個類別個別的 MMB 推論知識部份範例。

表 2　WSACS「B_1：職棒」類別的 MMB 推論知識範例

W_j	$p_{1\,1}$：$P(W_j \mid B_1)$	$\overline{p}_{1\,1}$：$P(W_j \mid \overline{B}_1)$
W_1：投手	0.91	0.25
W_2：捕手	0.88	0.22
……	……	……

表 3　WSACS「B_2：大學」類別的 MMB 推論知識範例

W_j	$p_{1\,2}$：$P(W_j \mid B_2)$	$\overline{p}_{1\,2}$：$P(W_j \mid \overline{B}_2)$
W_1：老師	0.92	0.32
W_2：學生	0.91	0.35
……	……	……

表 4　WSACS「B_3：醫院」類別的 MMB 推論知識範例

W_j	$p_{1\,3}$：$P(W_j \mid B_3)$	$\overline{p}_{1\,3}$：$P(W_j \mid \overline{B}_3)$
W_1：醫生	0.92	0.30
W_2：護士	0.88	0.25
……	……	……

＊推論引擎模組之範例

在 WSACS 開始推論動作之前，WSCS 會事先建構好第 H 層(假設 H=3)的網站分類目錄架構，如圖 5 所示。在建構好 WSACS 第 H 層的網站分類目錄架構之後，使用者便可以開始推論目標網站的類別隸屬。

假設網站分類人員現在要把「天下網路書店」這個目標網站進行分類的處理，則首先輸入天下網路書店網址列 http://www.cwbook.com.tw/cw/T1.jsp 到 WSACS 當中；而 WSACS 會運用網站結構探勘和網站內容探勘的方式(應用程式滲透)，將天下網路書店這個網站的所有內容網頁超連結架構和中文文字彙總成一個文件，然後交由 CKIP 斷詞器(斷詞單元)進行斷詞的動作，並進一步透過去除贅詞單元和同義詞決定單元的加值處理(去除重疊詞、一字詞和同義詞等贅詞)，進而產生如表 5 的網站詞集。

表 5　天下網路書店網站的網站詞集

詞編號	詞名	網站編號	出現次數
C0001	書名	B0001	72
C0008	定價	B0001	65
C0028	作者	B0001	81
C0045	考試	B0001	19
C0088	出版商	B0001	8
…	…	…	…

在求得天下網路書店的網站詞集後，WSACS 的推論引擎模組便會將這些詞集個別代入多目標類別的 MMB 推論公式當中，如表 6 所示。

表 6　天下網路書店網站 H=3 之多目標類別 p_{ij}、$\overline{p}_{ij}$ 值

類別編號	類別名	詞編號	詞名	p_{ij}	$\overline{p}_{ij}$
A0002	教學資源	C0001	書名	0.69	0.22
A0002	教學資源	C0028	作者	0.65	0.32
A0018	線上教學	C0001	書名	0.61	0.32
A0018	線上教學	C0028	作者	0.57	0.29
A0033	網路書店	C0001	書名	0.89	0.33
A0033	網路書店	C0028	作者	0.90	0.36
…	…	…	…	…	…

由表 6 可以清楚地看出在每一個類別當中，其所擁有的詞彙之 p_{ij}、$\overline{p}_{ij}$ 值大都不一致，也正因為如此，將天下網路書店網站的網站詞集代入多目標類別個別的 MMB 公式後，搭配每一個類別其所屬的 MMB 推論知識，將會產生高點網站書店網站屬於多目標類別個別的後天機率值，如表 7 所示。

表 7　天下網路書店網站隸屬於 H=3 之多目標類別個別機率值

類別編號	類別名	後天機率值	排名
A0002	網路書店	0.91	1
A0008	圖書資源	0.85	2
A0018	線上教學	0.69	3
…	…	…	…

WSACS 為了減少網站分類錯誤的情況發生，故設定候選類別其機率值的門檻至少要高於 0.8，而 WSACS 也會挑選機率值高於 0.8 的所有候選類別項，作為 WSACS 推薦給使用者 H=3 的類別可能項；其中，表 7 還會將所有候選類別的機率值由高至低排序，以避免當全部的候選類別項之機率值均低於門檻 0.8 時，WSACS 才能挑選機率值排名前 3 高的候選類別，作為 WSACS 推薦給使用者目標網站於 H=3 時的類別可能項的替代方案。

在 WSACS 產生了目標網站候選類別決策資訊集後，發現了 0.91-0.85=0.06；WSCS 假設機率值最高的候選類別項，減去機率值其次的候選類別項的數值若大於 0.05，則可直接選取機率值最高的候選類別項，作為目標網站最終的類別隸屬，而此例 0.06>0.05，故 WSACS 便會直接挑選「網路書店」這個候選類別決策資訊，作為高點網路書店網站最終的類別隸屬；但倘若今天機率值最高的候選類別項，減去機率值其次的候選類別項的數值若小於 0.05，則 WSACS 將會拿目標網站當中的網站詞集作為判斷的依據，也就是比較網站詞集個別在這些候選類別當中，其累積出現的次數最多，即為目

標網站最終的類別隸屬，而 WSACS 期望藉由這樣的方式，能夠秉持著用最客觀的方式來獲取最客觀的結果之願景。而在確認好目標網站的類別後，WSACS 會將目標網站當作一個新的知識訓練網站樣本，將目標網站的詞、p_{ij}、$\overline{p_{ij}}$ 值作知識學習的動作，進一步不斷地更新及修正 MMB 推論知識庫。

＊知識學習模組之範例

延續上述之範例，當 WSACS 確認「天下網路書店」網站的最終類別為「網路書店」時，WSACS 便將其視作為一個新的知識訓練網站樣本，並開始著手知識學習的動作，而表 8 為部分知識學習後 H=3 之多目標類別 p_{ij}、$\overline{p}_{ij}$ 值之範例。

表 8　知識學習後 H=3 之多目標類別的 p_{ij}、$\overline{p}_{ij}$ 值部分範例

類別編號	類別名	詞編號	詞名	p_{ij}	$\overline{p}_{ij}$
A0002	網路書店	C0001	書名	0.92	0.34
A0002	網路書店	C0028	作者	0.89	0.23
A0002	網路書店	C0352	拍賣	0.75	0.37
A0018	線上教學	C0001	書名	0.62	0.36
A0018	線上教學	C0028	作者	0.52	0.27
A0018	線上教學	C0352	拍賣	0.58	0.38
A0033	教學資源	C0001	書名	0.58	0.39
A0033	教學資源	C0028	作者	0.67	0.26
A0033	教學資源	C0352	拍賣	0.49	0.40
…	…	…	…	…	…

由表 8 可看出，加上底線的部分為經過知識學習模組的處理後所產生的變化。其中，WSACS 由天下網路書店網站當中學習到一個新的網站詞彙「拍賣」，故立即新增並計算其多目標類別個別的 p_{ij}、$\overline{p}_{ij}$ 值；其次，由於高點網路書店網站這個新的知識訓練網站樣本的加入，而又因為其網站類別已確定為「網路書店」，故在 WSACS 完成知識學習的動作後，「網站書店」類別的 p_{ij} 值會較其它非「網站書店」類別的 p_{ij} 值變動幅度大，相反地，非「網站書店」類別的 $\overline{p}_{ij}$ 值也會較「網站書店」類別的 $\overline{p}_{ij}$ 值變動幅度大，其原理如公式(1)和(2)所示。

(二)、WSACS 之系統運作流程介紹

此段落將以分段說明的方式來介紹和說明 WSACS 之運作流程，期望藉由這樣的方式讓讀者能夠更瞭解 WSACS 的真諦。

圖 6 為 WSACS 之系統主畫面，而 WSACS 的執行流程為，運用「知識建構」鍵來輸入網站的類別、名稱和網址，接著，執行四個步驟，然後，「顯示樣本資料」鍵可以快速地顯示樣本資料；運用「推論網站類別」鍵來推論目標網站類別，接著，執行四個步驟，然後，按「顯示後天機率值」鍵即可確認第 H 階層類別，且 WSACS 會在同一個時間內上推出目標網站的第 1 層到第 H-1 層；而「知識學習」鍵則可以快速地學習詞彙、p_{ij} 和 $\overline{p_{ij}}$ 值。

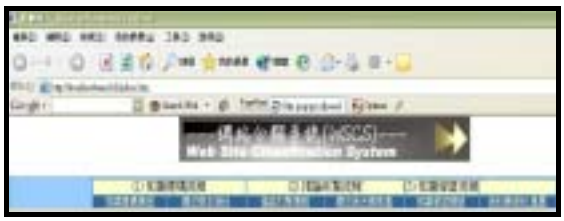

圖 6　WSACS 之系統主畫面

圖 7 為 WSACS 知識建構模組的第一步驟，使用者依序選擇知識訓練網站樣本的網站類別、輸入知識訓練網站樣本的名稱和網址，然後，按下「擷取」鍵，WSACS 即會開始探勘網站的超鏈結架構和網頁的中文文句。

圖 7　WSACS 知識建構模組的第一步驟

圖 8 為 WSACS 知識建構模組的第二步驟，使用者按下「斷詞程式」鍵可啟動 CKIP 斷詞器，接著按「下一步」。

圖 8 WSACS 知識建構模組的第二步驟

圖 9 為 WSACS 知識建構模組的第三個步驟，按下「斷詞程式」鍵，並選擇「開啟」鍵；使用 CKIP 斷詞器去開啟「ABC.txt」，並進行斷詞分析；並且按下「另存新檔」鍵，檔案要存在[WSACS]的目錄下，檔名為「DEF.txt」；結束後請關閉 CKIP 斷詞程式並按下[下一步]鍵。

圖 9 WSACS 知識建構模組的第三步驟

圖 10 為 WSACS 知識建構模組的第四個步驟，按下「開始計算」鍵，WSACS 即會開始運算網站詞集的 p_{ij} 和 $\overline{p_{ij}}$ 值，產生網站詞集更新後的 p_{ij} 和 $\overline{p_{ij}}$ 值，用以達到知識學習的目的。

圖 10 WSACS 知識建構模組的第四個步驟

圖 11 為顯示知識訓練網站樣本資料，使用者只要按下「顯示樣本資料」鍵，WSACS 即會顯示使用者所選定類別的知識訓練網站樣本集合的詳細資訊，包括「數量」、「擁有之詞彙數量」…等等。

圖 11 顯示知識訓練網站樣本資料

圖 12 為 WSACS 之推論引擎模組的第一個步驟，和知識建構模組的第一個步驟不同的是，使用者不需要指定目標網站的類別，WSACS 會以目標網站的辭彙為推論依據，推論出目標網站的類別隸屬。

圖 12 WSACS 推論引擎模組的第一個步驟

至於推論引擎模組的後三個步驟則和知識建構模組一致，故不再贅述。而圖 13 為 WSACS 推論目標網站類別隸屬的結果，使用者按下「顯示後天機率值」鍵，WSACS 即會推論出目標網站隸屬於多目標類別的個別機率值，並進一步挑選後天機率值最高的候選類別項，作為目標網站第 H 層(第 3 層)的正確類別項，同一時間內，WSACS 也會推論出目標網站的階層式(第 1 層到 3 層)類別隸屬。

圖 13 WSACS 推論網站類別隸屬的結果

在推論完目標網站的類別隸屬後，此一目標網站將會被視為一個新生的知識訓練網站樣本，並針對其辭彙、p_{ij} 和 $\overline{p_{ij}}$ 值做學習的動作，如圖 14 所示。

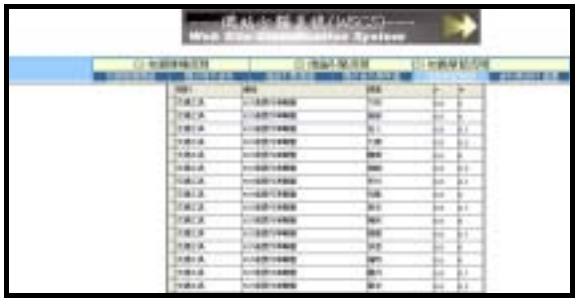

圖 14 WSACS 之知識學習模組的知識學習結果

(三)、各分類方式的比較表

本研究以 118 個知識訓練網站樣本，針對 WSACS 採用的 MMB 分類方式之實證結果，搭配嚴謹的人工[7]、不嚴謹的人工[7]、Bayesian[11]、向量[10]、Metadata[25]、Naive Bayes[13]和 Hybrid Method of Combining a Dictionary-Based Technique and a kNN Classifier[33]這七種的分類方式之研究結果，本研究以「準確率」、「範圍限制」、「花費時間」、「複雜度」和「客觀度」這五個評比項目，彙整了一個簡易的比較表，至於詳細的數據資料如表 9 所示。

表 9　MMB 分類方式與其他分類方式的比較表

評比項目 / 分類方式	準確率	範圍限制	花費時間	複雜度	客觀度
[1]MMB(2005)	0.88	無	中	低	高
[2]嚴謹的人工	0.90	無	長	高	高
[3]不嚴謹的人工	0.70	無	中	高	中
[4]Bayesian(2003)	0.80	有	短	低	低
[5]向量(2002)	0.85	無	長	高	中
[6]Metadata(2001)	0.85	有	短	中	低
[7]Naïve Bayes(2001)	0.66	有	短	中	低
[8]Hybrid(2003)	0.77	有	短	中	中

由表一可看出，採用 MMB 來執行網站分類的動作，不僅保有 88%的準確率，在「範圍限制」、「複雜度」和「客觀度」都較其他的分類方式佳，目前 WSACS 仍有改進空間的地方即為「花費時間」，故未來 WSACS 會致力於訂定更完善的規則，以便能夠達成「以最少的推論知識，即能推論出最正確的網站類別」之願景。

五、未來展望

目前網站大多採單一主題分類(Multi-class)，但有許多的網站所討論內容屬於多主題，因此在分類上可能會產生錯誤現象，未來我們希望能作多主題類別(Multi-label, Multi-class)的研究，以提昇網站分類準確率。

至於 WSACS 未來的展望大致上還有以下三點，詳細的敘述和說明如下所示：

(一)、尋找更有效的「新增類別」方式

由於目前 WSACS 之網站分類目錄架構是綜合各大入口網之網站分類目錄所建構而成，然而，隨著時代的演進，未來 WSACS 仍然需要不斷有效地新增網站類別項目，以維持網站分類的正確性，否則倘若 WSACS 持續將一些較難辨識類別的網站放置在「其他」的類別當中，這樣的做法將會大幅降低 WSACS 的價值。有鑑於上述之考量，WSACS 未來將嘗試針對被歸類到其它類別當中的知識訓練網站樣本，從中擷取具代表性的關鍵字以作為新類別的規則所在，而至於新類別的名稱，則將會以詞頻前 5 高的關鍵字為依據來命名。

(二)、有效地拉大「誤差門檻值」

由於 WSACS 所預設的誤差門檻值為 0.05，這代表後天機率值集當中，倘若 WSACS 計算出之機率值第 1 高的候選類別項和其它的機率值集之差距大於 0.05，WSACS 便會直接認定機率值第 1 高的候選類別項為目標網站最終的類別隸屬；反之，倘若 WSACS 計算出之機率值第 1 高的候選類別項和其它的機率值集之差距小於 0.05，WSACS 未來的工作，便是使用一個有效的方法將誤差值拉大，讓分類結果更具有說服力。

(三)、挑選網站鏈結結構「最底層」的網頁群組作為推論知識之訓練範圍

有鑒於 WSACS 目前蒐集推論知識的範圍，為網站內每一階層的網頁內容中文文句，故在處理時間上將會耗去太多的時間，然而，在我們的研究當中發現，網站當中含有一些和網站類別隸屬毫無關聯的雜質網頁，分散在中間階層的網頁群組當中。也因此，未來 WSACS 將會將「第一層」、「中間層」、「最底層」和「綜合每一層」這四種方案做比較，試圖去找出哪一種方案所擷取出的推論知識較具代表性。在一些簡單的實驗中，我們赫然發現最底層的網頁群組當中的辭彙較具代表性，未來 WSACS 將會採集更多的樣本來驗證此一假說，倘若此假說成立，則 WSACS 的處理時間之效能將會大幅提昇。

(四)、將「網頁鏈結」列入判斷網站類別隸屬的因子「網頁鏈結」列入

近來已經有愈來愈多有關網站自動分類的技術文章，將「網頁鏈結」列入判斷網站類別隸屬的因子，期望藉由這樣的方式來提昇網站自動分類的準確率和客觀度，未來 WSACS 也將會致力於探討「網頁鏈結」是否真的會影響網站自動分類的結果。

(五)、可應用於「電子市集」

現今社會由於科技的持續進步，傳統的面對面交易模式已演進成現今藉由網際網路來交易的電子商務模式；而傳統上顧客必須到一般商店或市場才能購買到的商品，現今也演變成為藉由電子市集，顧客即可透過網際網路購買商品。然而，現今網站市集最為人詬病的地方，不外乎是缺乏一套有效的網站類別群組搜尋機制，以便利提昇網站市集當中買賣雙方網站群組之品質，進而提昇顧客使用其網站市集來購買或販賣商品的意願。

未來 WSACS 將可以採用多重代理人(multi-agent)的機制，建構三個智慧代理人(intelligence agent)，分別為買方、賣方和公正第三者三種類型的網站群組，透過 WSACS，使用者可以快速地建構好這三種類型的網站群組，藉此來建構一個品質良好的電子市集平台。且由於 WSACS 可以做多目標類別的分類，故屆時運用 WSACS 所建構的電子市集，其所屬的類型便會趨於多元化，而不會有範圍上的限制，且由於 WSACS 之網站分類的準確率很高，故確保了電子市集平台當中買方、賣方和公正第三者，其網站群組的正確性，進而提供了使用者一個建立品質良好的電子市集平台之工具。

六、結論

隨著網際網路的日新月異，網站的數量將會呈倍數成長，也因此運用有效的網站自動分類技術，將可替代繁瑣又需要耗費大量人工處理的資料分類作業，進而大幅地改進網站管理的效率。而本研究進一步彙總中文網站自動分類技術能帶給應用者的好處有，「強化網站管理的效率」、「增進知識管理的效能」和「大幅減少人力、時間和成本」三點。為了要有效地獲取上述的三個好處，本研究提出了三個提昇效能方法，分別為「去除多餘的推論知識」、「將推論知識最佳化」和「不斷地作知識學習的動作」，期望藉由這樣的方式來提昇 WSACS「分類準確率」和「系統處理時間」的效能。

目前 WSACS 在自動辨識網站類別的準確率之效能已達到 88%的高水準，明顯較其他的網站自動分類方式來的好，然而，系統處理時間較長的問題則是目前 WSACS 迫切需要解決的地方，也因此未來 WSACS 將會把重心放在提昇「系統處理時間」的效能。

七、誌謝

感謝中央研究院的中文知識庫小組，免費提供 CKIP 中文斷詞程式給本研究作為學術上的研究根基，而斷詞程式的網址來源為 http://ckip.iis.sinica.edu.tw/CKIP/ws/。

八、參考文獻

[1] 李中彥、駱思安、吳宏文，“MMB 中文網站階層式分類推論知識之建構”， 2004 年國際資訊管理暨電子商務經營管理研討會，光碟論文集(場次：3-3、論文發表編號：IN3-6、嘉義：南華大學)，2004.

[2] 李中彥、駱思安、林佑威，“網站分類系統推論知識品質之提昇”，ICIM2005 第十六屆國際資訊管理學術研討會，光碟論文集(場次：session A、場地：第 10 研討室、台北：輔仁大學)，2005.

[3] 李中彥、駱思安、黃如盛，“應用 MMB 建構中文網站階層式分類推論引擎”，2005 年學習型知識社群與電子商務實務研討會，光碟論文集(場次：2-1、場地：國際會議廳、論文接受編號：2-1-2、台北：中國文化大學)，2005.

[4] 李中彥、駱思安，“以 Web Services 建構網站分類推論系統”, 2005ING 安泰管理碩士論文獎暨研討會，光碟論文集(場次：411、類別：資訊管理 7A 佳作、台北：台灣科技大學)，2005.

[5] 吳國榮，“階層式類別架構的學習於文件分類之探討研究”，中正大學資訊工程研究所，碩士論文，2000.

[6] 吳宜鴻，“全球資訊網資料之分析、索引與擷取”，清華大學資訊工程研究所，博士論文，2000.

[7] 邱志宏，“個人網路資訊管理系統及其網頁分類方法之研究”，銘傳大學資訊管理研究所，碩士論文，2002.

[8] 唐大任，“中文斷詞器之研究”，交通大學電信工程研究所，碩士論文，2001.

[9] 張啟峰，“整合階層式分類目錄的演算法設計及評估”，中正大學資訊工程研究所，碩士論文，2001.

[10] 曾耀順，“在超連結環境下針對資訊分類相關權威網頁之探勘”，成功大學電機工程研究所，碩士論文，2002.

[11] 游佳琪，“網站類別辨識推論系統及知識管理”，中國文化大學資訊管理研究所，碩士論文，2004.

[12] 駱思安，”以 Web Services 建構中文網站階層式分類推論系統”, 中國文化大學資訊管理研究所，碩士論文，2005.

[13] A. McCallum et al., “A Machine Learning Approach to Building Domain-Specific Search Engines”, in proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), 1999, pp. 662–667.

[14] C.-Y., Lee、C.-C., Yu, “Decision on Classifying Chinese Commercial Web Sites by Bayesian Approach,” in proceeding of the Fourth Annual Hawaii International Conference on Business, 2004.

[15] C.-Y., Lee、Evens, M.、Carmony, L.、 Trace, D. A.、Naeymi-Rad, F., “Recommending Tests in a Multimembership Bayesian Diagnostic Expert

System," in proceedings of the Fourth Annual IEEE Symposium on Computer Based Medical Systems, 1991, pp. 28-35.

[16] C.-Y., Lee, "On Using Bayesian Approach Recognizing Chinese Electronic Bookstore Web Sites", in proceeding of the Tenth ISSAT International Conference (Reliability and Quality in Design), 2004, pp. 290-294.

[17] G. Siolas、F. d'Alché, "Support Vector Machines based on a Semantic Kernel for Text Categorization", in proceedings of the IEEE-INNS-ENNS International Joint Conference on, 2000, pp.205-209.

[18] Jyh-Jong Tsay、Jing-Doo Wang, "Improving Automatic Chinese Text Categorization by Error Correction", in proceedings of Information Retrieval of Asian Languages(IRAL '00), 2000.

[19] K.-J., Chen、S.-H., Liu, "Word Identification for Mandarin Chinese Sentences", in proceedings of COLING92, 1992, pp.101-107.

[20] Leah S. Larkey、W. Bruce Croft, "Combining Classifiers in Text Categorization," in proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, , 1996, pp.289-297.

[21] Mladenic, D.、Institute , J.S., "Text Learning and Related Intelligent Agent：A Survey", in proceedings of IEEE Intelligent Systems, Intelligent Information Retrieval ,1999, pp.44-54.

[22] M. Sasaki 、K. Kita, "Rule-Based Text Categorization Using Hierarchical Categories," in Proceedings of SMC-98, IEEE International Conference on Systems, Man, and Cybernetics, 1998.

[23] Oh-Woog Kwon、Jong-Hyeok Lee, "Web Page Classification Based on k-Nearest Neighbor Approach", in proceedings of the 5fth international workshop on Information retrieval, 2000, pp.9-15.

[24] Richardo Baeza-Yates、Berthier Ribeiro-Neto, "Modern Information Retrieval", Addision Wesley Longman Limited, 1999.

[25] R. Ghani、S. Slattery 、Y. Yang, "Hypertext Categorization Using Hyperlink Patterns and Metadata," in proceedings of ICML-01, 18 th International Conference on Machine Learning, 2001.

[26] S. Dumais、H. Chen, "Hierarchical Classification of Web Content," in proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000, pp.256-263.

[27] T. Joachims, "A Statistical Learning Model of Text Classification for Support Vector Machines," in proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp.128-136.

[28] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", in proceedings of 10th European ECML Conference on Machine Learning, 1998, pp.137-142.

[29] V. Vapnik, The Nature of Statiscal Learning Theory. Springer, New York, 1995.

[30] W. Lam、C. Y. Ho, "Using A Generalized Instance Set for Automatic Text Categorization," in proceedings of the 21th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp.81-89.

[31] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization", 1999.

[32] Y. Yang、X. Liu, "A re-examination of text categorization methods," in proceedings of the 22th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp.42-49.

[33] Y.-M., Chang、Y.-H, Noh, "Developing a specialized directory system by automatically classifying Web documents", in proceedings of journal of information science", 29 (2) 2003, pp. 117–126.