行政院國家科學委員會專題研究計畫成果報告
電子報系統中個人化技術的設計與實作(II)
Design and Implementation of Personalization Technology in E-News Systems (II)

計畫編號：NSC 90-2213-E-004-003
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：沈錳坤　政治大學資訊科學系

一、中文摘要

由於全球資訊網的興起，帶動了電子報的發展。然而隨著電子報的普及，資訊量愈來愈多。如果能將個人化的技術應用在電子報上，將會為閱聽人帶來相當大的便利性。個人化階層分類建立符合閱聽人閱讀習慣的分類，以方便閱聽人閱讀新聞。我們利用中文關鍵詞擷取、關鍵詞篩選及個人化分類階層建造法為閱聽人造出個人新聞分類階層，讓閱聽人只看到他感興趣的新聞。除此之外，利用新聞的分類規則，我們亦做到個人化的「新聞內容」分類的功能。因此我們實做出 News Boy 個人化電子報系統。在第一年的計畫中，News Boy 提供個人化排列及社群推薦等功能。第二年的計畫則進而將發展重點擺在個人化階層分類及個人化新聞內容分類兩部分。

關鍵詞：電子報、個人化、個人化分類、個人化內容分類

Abstract

Owing to the prevalence of the Internet and WWW, many kinds of web sites are built up. And electric newspaper is one of them. The best way is to use the power of science and to add personalization to electric newspaper. A personalized electric newspaper can only show the news the reader wants and can distribute news to the layout the reader expects.

Besides, we implement a personalized electric newspaper system, News Boy (http://newsboy.cs.nccu.edu.tw). In the project of the last year, Newsboy supports the functions of personalized ranking, recommendation, and preference learning. In this year, we emphasizes on the functions of personalized categorization and personalized content classification.

Keywords: E-News, Personalization, Personalized Categorization, Personalized Content Classification

二、緣由與目的

由於 Internet 和 World Wide Web 的盛行，五花八門的網站如雨後春筍般的成立，電子報也是其中之一。然而，單單只是把平面的報紙移植到網路上，或是藉 HTTP 的便利結合 multimedia 而成的電子報，只是報紙、廣播及電視的結合而已。唯有利用科技的力量，加入個人化的能力，才能讓電子報真正的獨樹一格。因為電視節目不能只播報閱聽人感興趣的新聞；報紙不能為閱聽人把新聞分到閱聽人預期的版面，可是個人化電子報，卻可以做的到上述的幾項，是一個真正為每一個閱聽人量身打造的個人化電子報系統。

以現在一天所發生的新聞數量來看，少說都有二、三百則，而以「人」的角度來看，二、三百則新聞以線性的方式一則則呈現，對閱聽人來說，是一件很大的負擔，更遑論說要從其中去找出同類的新聞了。所以個人化階層分類就是再進一步將新聞分類，並以階層的方式呈現，且加上個人化，讓閱聽人都具有各自的個人新聞分類階層，可以只看到他感興趣的新聞。換言之，如果以報紙版面來舉例的話，每個人都有自己的頭版。例如王先生想要在台北縣置產，則他的頭版可能為「房地產」下面的「台北縣」。在個人化新聞分類階層裡面，我們先對分類階層作個人化，再對各分類裡的新聞作個人化排列。而個人化新聞內容分類，則是將一則新聞分到閱聽人預期的那一個分類裡，例如一則有關「馬英九參加明星籃球賽」的新聞，對「政治」感興趣的閱聽人來說，可能屬於「政治」，對「娛樂」感興趣的閱聽人來說，可能屬於「娛樂」，而對「運動」感興趣的閱聽人來說，可能屬於「運動」等。我們並且實做出 News-Boy 個人化電子報系統(http://newsboy.cs.nccu.edu.tw)，提供了個人化排列(Personalized Ranking)，社群推薦(Recommendation)，自動新聞收集和自動學習閱聽人喜好(Preference Learning)，以及個人化階層分類(Personalized Categorization)，個人化新聞內容分類(Personalized Content Classification)等功能。第一代的 News Boy 實做出了個人化排列及社群推薦等功能。第二代的 News-Boy 則將發展重點擺在個人化階層分類及個人化新聞內容分類兩部分。

三、結果與討論

1 原理

我們的個人化階層分類，所會運用到的技術包括中文關鍵詞擷取、關鍵詞篩選，以及後面會提到

的個人化分類階層建造法。

個人化階層分類將閱聽人的喜好以樹狀階層表示，閱聽人的喜好改變相當於調整閱聽人的個人新聞分類階層，所以我們先利用個人化分類階層建造法建造出階層，再利用中文關鍵詞擷取出一則新聞的屬性，和每個分類所具有的 keyword-list 相比較，以決定該分到哪一個分類。等到閱聽人閱讀完新聞之後，再依據閱聽人的喜好，調整分類階層，達到個人化的目的。

個人化新聞內容分類則是在閱聽人已經具有個人化新聞分類階層之後，依據新聞分類規則，將新聞分到閱聽人預期的新聞分類去。

1.1 中文關鍵詞擷取

要判斷出一則新聞的屬性，必須由它所包含的文字詞彙判斷起。目前主要有三種關鍵詞擷取的方法，詞庫比對法，文法剖析法及統計分析法。[2]第一種為詞庫比對法：從已經有的詞庫中，一一的比對關鍵詞有否出現在輸入的文句中。這種比對法不需要自然語言的技術，可是無法擷取出在該詞庫未出現的關鍵詞，例如現代的人名或新設立的機關名。第二種為文法剖析法：利用自然語言處理技術的文法剖析程式，來剖析出文件中的名詞片語。由於因為運用到自然語言處理技術，因此如果是出現在標題，書目等的單獨不成句的關鍵詞就無法擷取出來。第三種方法為統計分析法：經由分析文件，計算每一個字詞出現的頻率，取出大於某一頻率的字詞，就是關鍵詞。但是因為沒有參考辭庫，所以可能發生取出無意義的字詞。可是卻可以利用來擷取出辭庫所未收集的專業用語。News Boy 所使用的是詞庫比對法，而所用的詞庫為蔡志浩先生所建置的詞庫(http://www.geocities.com/hao510/wordlist/tsaiword.zip)，約有 12 萬多個詞。

然而並非所有擷取出來的關鍵詞，都具有代表性。因此我們必須先把無用的贅詞去掉，留下真正具有代表性的關鍵詞來代表該則新聞。一般最常用的兩個判斷依據為 TF(Term Frequency) 及 DF(Inverse Document Frequency)。TF 指的則是一個關鍵詞在一則新聞裡的出現頻率。DF 指的是一個關鍵詞在所有新聞裡面的出現頻率。而判斷一個關鍵詞是否為重要關鍵詞，是否具有代表性，就是利用這兩個值再依據判斷公式來決定。最簡單的就直接利用這兩個的比值來決定，TF/DF。我們採取的是 Entropy weighting[3]的方式來決定，因為這個方法具有較好的效果。公式如下：

$$a_{ik} = \log(TF_{ik} + 1.0) * \left(1 + \frac{1}{\log(N)} \sum_{j=1}^{N} \left[\frac{TF_{ij}}{DF_i} \log\left(\frac{TF_{ij}}{DF_i}\right) \right]\right)$$

a_{ik} 即為用來判斷關鍵詞 i 是否為 document k 的重要關鍵詞的數值，N 為所有 document 的數量。我們可以預先給定一個值，當 a_{ik} 大於該值，我們才判斷其為具代表性的關鍵詞。

1.2 個人化分類階層建造法

個人化階層分類的精要在於，每個閱聽人擁有自己的分類階層，此分類階層隨閱聽人的喜好而調整，也就是說，此分類階層即可代表閱聽人的喜好。我們將此階層稱為個人化分類階層。接下來說明個人化分類階層的結構，以及我們如何造出個人化分類階層。

分類階層的資料結構

閱聽人的分類階層是以一棵 ordered-tree 來表示，所謂的 ordered-tree 是指在同一個 parent 下的 children，具有順序性。我們先做下列定義：

Node：包含 Conceptual node 及 Term node 兩部分。

Conceptual node：我們針對所有的新聞所定出的最初的可能分類種類。主要來自於一般常見的分類集合，例如搜尋引擎等。事先已經決定，和新聞內容無關。只能出現在 Class level 裡，可能的 children 有 Conceptual node、Term node、News。

Term node：由新聞內容所提取出來的重要關鍵詞組成。由輸入的新聞來決定，依據輸入的新聞不同而不同。可能出現在 Class level 裡及 News level 的最上層。可能的 children 為 News。

News：新聞。

Conceptual node set：由所有的 Conceptual node 所構成。

Root：整個階層的最上面。

Class level：階層的最上層部分。不包含 Root。可能包含 Conceptual node 和 Term node。

News level：階層裡除了 Class level 外的部分。不包含 Root。可能包含 Term node 和 News。

個人化分類階層主要由三個部分所組成，最上面的 Root，以及接下來的 Class level 和 News level。Class level 為閱聽人的喜好階層，也就是能代表閱聽人喜好的分類階層。News level 則依據每次輸入的新聞不同而變。整個階層如圖 1。

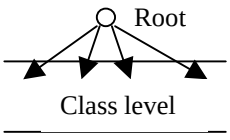

圖 1：整個階層結構

接著我們就上面我們所提到各種定義和我們所做的個人化分類階層關係做個說明。上面所提到的 Node，就相當於階層裡的新聞的索引，對於所有輸入的新聞，我們依照 Node 將之做分類，Node 和 News 的關係如圖 2。先對所有的新聞依照 Class level 一層層做分類，再針對每一個分類下的新聞自己產生 Term nodes，如此造出整個階層，省略 News 部分後如圖 3。

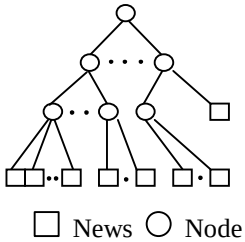

圖 2：Node 和 News 在階層裡的關係

另外，新聞具有所謂的熱潮的特性，常常突然發生的一些事，就變成大眾討論的焦點；而過一段

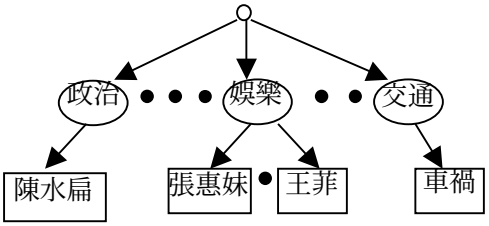

圖 3： ○ Conceptual node
□ Term node

時間之後，當大家的注意力開始移到別的地方的時候，它就不再是新聞的焦點了。正因為新聞具有這些特性，所以如果我們僅僅用 Conceptual node 來表示的話，沒有辦法表示出新聞的特性，所以我們提出 Term node，就是為了表示出上述的新聞的特性。

當某些事成為新聞的焦點時，自然會有相當多的新聞報導相關的事情，所以具有代表性的相關的詞彙就會在許多的新聞裡面重複的出現，這些具有新聞焦點代表性的相關詞彙，就會被我們擷取出來成為 Term node，當閱聽人對該類的新聞有興趣時，則該 Term node 就會變成閱聽人的喜好之一，如果再有相關的新聞，閱聽人都可以很快的看到；而一段時間過去之後，當它的新聞熱潮消失之後，或者閱聽人對之再不感興趣時，那它自然會從閱聽人的喜好中被刪除。如此符合了新聞的特性。

分類階層的基本操作

如前所述，分類階層可代表閱聽人的喜好，所以一旦閱聽人的喜好變動，也就等於閱聽人的分類階層的調整。階層的調整也就是階層中的 Node 的移動與增刪，所以我們定義了 Node Operation。另外在我們的個人化階層裡面，閱聽人的喜好順序表示為，同一個 parent 下的 child 裡，左邊大於右邊。

Node Operation：針對 Node 的操作，如下所列：

・Up-level：往上移到上一層。如圖 4 的 C。
・Down-level：往下移到下一層。如圖 4 的 D。
・Move-Right：同一層向右移一位。如圖 4 的 B。
・Move-Left：同一層向左移一位。如圖 4 的 A。
・Insert：插入階層中。如圖 4 的 F。
・Delete：從階層中移除。如圖 4 的 E。

其中 Move-Right、Move-Left 只能作用在同一個 parent 下。

接下來我們以 Conceptual node 及 Term node 為主體來討論其相關的 Operations。

Conceptual node：

・Up-level：最多到達 Root 之下的第一層。
・Down-level：最低只能到原本所屬的 level。
・Move-Right：限於同一個 parent 之下。
・Move-Left：限於同一個 parent 之下。
・Insert：如果該 Conceptual node 不存在於階層中時，才可使用。而且只能 insert 在原本所屬的 parent 下。
・Delete：如果該 Conceptual node 存在於階層中時。

Term node：

・Up-level：最高只能到達 Root 之下的第一層。
・Down-level：最低只能到 Class level 的最底層。
・Move-Right：限於同一個 parent 之下。
・Move-Left：限於同一個 parent 之下。
・Delete：如果該 Term node 存在於階層中時。

對於關鍵詞而言，當它有足夠的重要性被選為 Term node 時，就會出現在 News level 的最上層，成為整個階層的一部份了，而當它被 Delete 之後，除非它重新出現在新聞裡面，並再次具有足夠的重要性而成為 Term node，否則是不會有再回到階層中的機會。所以對 Term node 而言，是不存在 Insert 的。

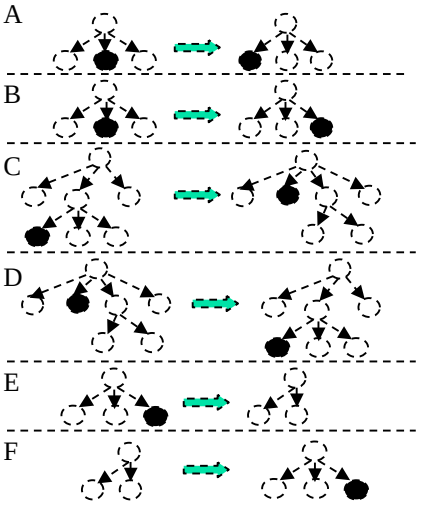

圖 4：Node 的相關 Operation

分類階層的個人化

關於個人化的部分，是對階層裡的 Node 做相關的 Operations，以達到個人化的目的，而 Node 的 Operation 則是以整棵此 Node 為 r o o t 的 Subtree 為單位。

另外，如果有 Node 做 Up-level 時，代表閱聽人對該 Node 的喜好大於它的 parent，所以一旦做 Up-level，將該 Node 放置在與它的 parent 同一個 parent 下，因為由左而右有順序性，所以放到它的 parent 的左邊。

實際操作如下：

Initial：因為尚未閱讀過任何新聞，所以 Class level 為一開始預設的狀態，每個 Conceptual node 都出現在階層預定的位置中，而 News level 則依照輸入的新聞和輸入新聞裡所找出的 Term node 組成。

隨著時間過去，閱聽人的喜好逐漸明顯：

當此 Node 符合閱聽人的喜好時，對此 Node 做相關的 Operations：

1.Conceptual node：

・Move-Left：如果此 node 的左邊尚有其他 Node。
・Up-level：如果此 node 的左邊沒有其他 Node。
・Insert：當此 node 不存在階層中。

2.Term node：

・Move-Left：如果此 node 的左邊尚有其他 Node。
・Up-level：如果此 node 的左邊沒有其他 Node。

當此 Node 不符合閱聽人喜好時，Conceptual node 及 Term node 的 Operations 都相同如下：

・Move-Right：如果右邊尚有其他 Node 時。
・Down-level：如果右邊沒有其他 Node 時。
・Delete：當位於原本所屬 level 的 parent 的所有 children 的最右邊，且閱聽人的對之不感興趣時。

到最後就變成為閱聽人量身打造的個人化階層了。

以下舉實例說明個人化階層的建造及對 Node

的操作，我們省略 News 的部分，以方便解釋 Node 的操作。為了方便說明，假設每次造出來的 Term node 都一樣。

圖 5 為閱聽人初次使用階層的狀態，整個階層的 Conceptual node 都還存在：

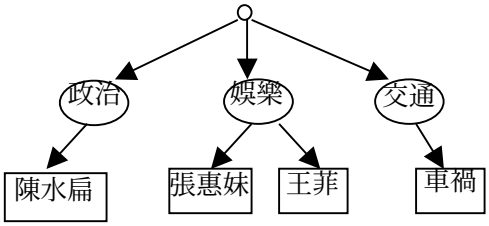

圖 5：階層一開始的初始狀態。

隨著時間的過去，發現閱聽人對「娛樂」的喜好比「政治」大，所以對 Conceptual node「娛樂」做 Move-Left。如圖 6。

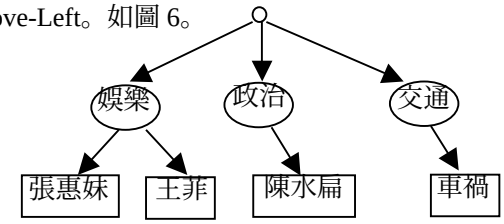

圖 6： 「娛樂」移到「政治」的左邊

發現閱聽人對於「張惠妹」的喜好很大，所以對 Term node「張惠妹」做 Up-level。如圖 7。

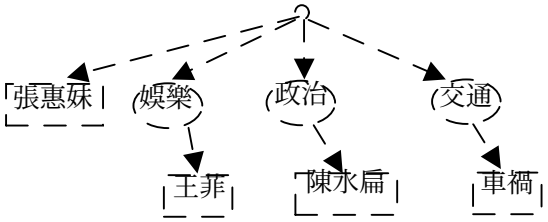

圖 7： 「張惠妹」Up-level

閱聽人對於「交通」沒有什麼興趣，所以「交通」被 Delete 了。如圖 8。

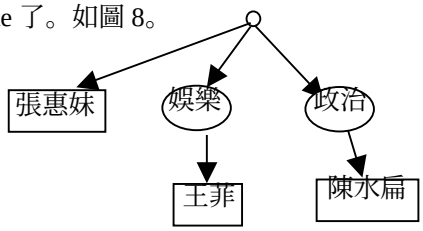

圖 8： 「交通」被 Delete 了

隨著每個人不同的喜好，再加上對 Node 的 Operations，自然而然就會為閱聽人造出獨一無二的個人化分類階層。

個人化新聞內容分類

有了閱聽人的喜好階層之後，接下來我們為每一個 Conceptual node 定義了一個 keyword-list，負責描述該 Conceptual node 的屬性，就如同每一則新聞有其 profile 描述其屬性一般，利用同樣的方法，只要把一則新聞的 profile 和 Conceptual node 的 keyword-list 做內積之後所得的值，作為相似度，比較一則新聞和該層的 Conceptual node 的相似度即可決定該分到哪一類。分類規則如下：

1. 如果該層的 Node 裡面，包含 Term node，且該則新聞包含此 Term node 所代表的關鍵詞，則直接分到 Term node 的分類去。若有多個 Term node 符合此條件，則分到較左邊的 Term node 去。
2. 如果該層的 Node 裡面沒有包含 Term node，或者是該則新聞裡面沒有具有 Term node 所包含的關鍵詞，則此則新聞就分到相似度最高的那一類去
3. 如果相似度一樣時，則分到喜好較高的那一類。

藉由分類規則的第三條規則，我們可以實做出「個人化新聞內容分類」，也就是能以閱聽人的角度來為新聞作分類。舉例來說，「馬英九參加明星籃球賽」的新聞，可能同時具備有「政治」、「影視」、「運動」等可能分類，我們利用新聞分類的第三條規則，就可以做到依照閱聽人的角度去自動幫他分到他認為的那一類去。

閱聽人喜好的學習

關於閱聽人喜好的學習，我們反應在依據閱聽人對每個分類的喜好程度來調整閱聽人的個人化階層上。而閱聽人對於每個分類的喜好程度，來自於閱聽人閱讀該分類下的新聞，大致可以以下兩者為判斷依據：

1. 閱讀順序：閱聽人對於各分類新聞的閱讀順序，越先讀的分類，表示越有興趣。
2. 閱讀新聞比重：在一個分類下的所有新聞裡，閱聽人閱讀的比例，比例越高表示越有興趣。

調整閱聽人的個人化階層的方法為以閱聽人的「閱讀順序」為主，而以「閱讀新聞比重」輔。至於閱聽人對於階層中每一個分類的「閱讀新聞比重」，我們以點閱新聞的數量來決定。而 parent 分類的「閱讀新聞比重」，我們以它所有 children 的「閱讀新聞比重」的總和，再除以 1/2，來當作它所具有的「閱讀新聞比重」，如此一來可以確保除非閱聽人真的對該分類下的某個 child 情有獨鍾，才對該 child 作 Up-Level。

2 系統架構

圖 9 為 News boy 的系統架構圖，主要分為 Server 及 Client 兩部分。以下將分別敘述如下

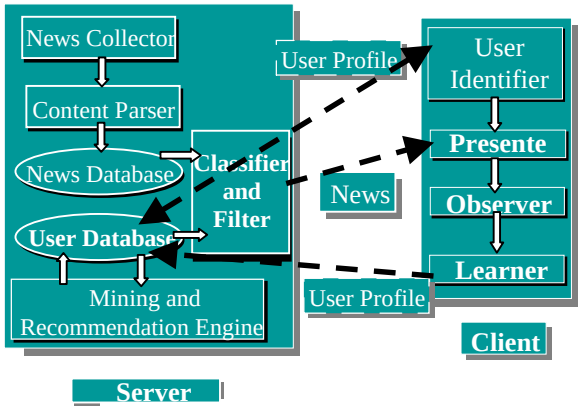

圖 9： News boy 系統架構圖

2.1 Server

1. 1.News Collector：負責 News Boy 的新聞收集工作，把收集到的新聞傳到 Content Parser。
2. 2.Content Parser：利用中文關鍵詞擷取技術，擷取從 News Collector 傳過來的新聞中的關鍵詞，再連同該新聞的報紙名稱，標題，系統時間，內文及擷取出來的關鍵詞都一併加入 News Database 之中。

3. Classifier and Filter：將 News Database 裡閱聽人尚未閱讀過的新聞分類，再針對每一類利用 content based filtering 進行個人化排列，最後把結果傳到 Client 端的 Presenter。
4. News Database：在我們的 News Database 中，會把下列與新聞相關的資訊存入。
5. User Database：存放 User 的個人相關資訊。
6. Mining and Recommendation Engine：定期的將閱聽人分出數個的社群，並找出其在閱讀時的共同特點，並各別加以推薦。

2.2 Client

1. User Identifier：負責辨識這個連上線的閱聽人是否來過本系統。
2. Presenter：依據閱聽人的個人喜好呈現新聞，在呈現的同時，啟動 Observer，觀察閱聽人。
3. Observer：觀察閱聽人的喜好，交給 Learner。
4. Learner：依據 Observer 的觀察，將閱聽人的相關資訊加以更新，再將結果送回 server 端的 User Database 儲存。

2.3 系統運作流程

當閱聽人通過身份認定之後，Server 端會把閱聽人尚未看過的新聞和閱聽人存於 User Database 的 Class level 傳送到 Classifier and Filter，先依 Class level 將新聞分類，再由新聞中擷取出 Term node 合成整個階層，接著將此階層傳送到 Client 端的 Presenter，再以階層的方式呈現給閱聽人，並於此同時啟動 Observer 觀察閱聽人的動向，並將觀察所得的喜好傳送給 Learner，由 Learner 更新閱聽人喜好，並於閱聽人結束使用時回傳到 Server 端的 User Database 儲存。

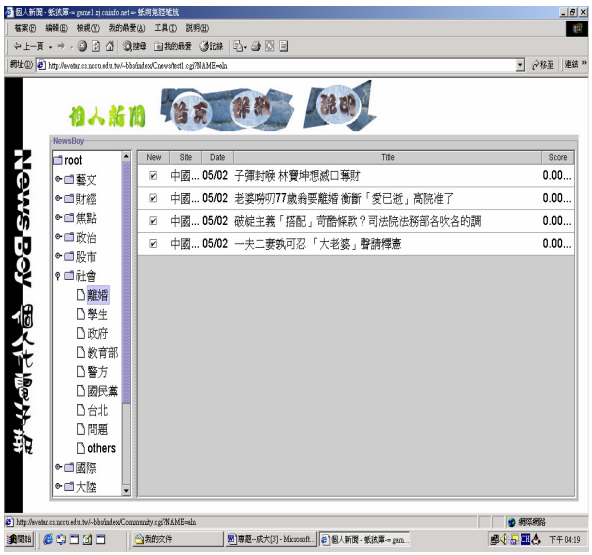

圖 10:中文階層分類

2.4 系統畫面

News Boy 除了具有中文新聞的分類階層(如圖 10)，亦有以 CNN(http://www.cnn.com)的新聞，做英文的關鍵詞擷取、關鍵詞篩選及個人化分類階層建造法，提供英文的個人化階層 (如圖 11)。

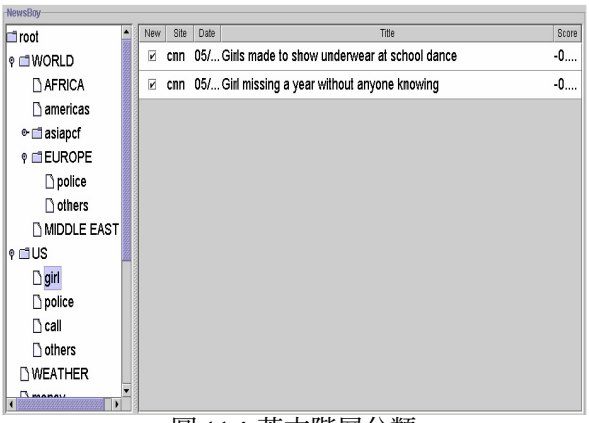

圖 11：英文階層分類

四、計畫成果自評

本研究計畫延伸第一年所發展的個人化電子報系統 Newsboy，Newsboy 新增的功能包括

1. 個人化階層分類 (Personalized Categorization) 。
2. 個人化新聞內容分類(Personalized Content Classification) 。

本研究的部分成果已經發表在 2002 Symposium on Digital Life and Internet Technologies。此外部分的成果也準備投稿至 2003 Pacific-Asia Conference on Knowledge Discovery and Data Mining。

五、參考文獻

[1] Kaoru Kobayashi, Yasuyuki Sumi, and Kenji Mase, Information Presentation Based on Individual User Interests. Proceedings of 1998 Second International Conference on Knowledge-Based Intelligent Electronic System, IEEE, April, 1998.

[2] Berners-Lee, R. Fielding, and H. Frystyk, Hypertext Transfer Protocol - HTTP/1.0. RFC 1945, May, 1996.

[3] Hidekazu Sakagami and Tomonari Kamba, Learning Personal Preferences on Online Newspaper Articles from User Behaviors. Proceedings of the 1997 Sixth International World Wide Web Conference, Senta Clara, CA, April, 1997.

[4] Philip S. Yu, Data Mining and Personalization Technologies. Proceedings of International Conference on Database Systems for Advance Applications, HsinChu, Taiwan, ROC,1999.

[5] 曾元顯，關鍵詞自動擷取技術之探討，中國圖書館學會會訊 106 期，九月，1997。