題目：個人化知識表徵瀏覽模型

摘要

電子資源相關檢索技術的發展，加速資源數位化的腳步，現今網路上已存在為數龐大的數位化資源。使用者可以透過搜尋引擎來檢索所需之資源。但隨著資訊爆炸性的成長與更新，面對如此巨大的網路虛擬資料庫，人們更難以從中截取或組織所截取的資源。單純的搜尋網頁或電子資源，已無法滿足使用者的資訊需求。面對如此龐大而豐富的資料，若能提供個人化的資訊搜尋與運用，使用者將能快速的掌握所需決策訊息，更可為個人創造最佳的成功契機。

藉由網路上所提供的搜尋引擎，我們可以很輕易從網路上搜尋到所需的資訊，再透過瀏覽器來點閱搜尋到的資料，相當的容易且方便，但這樣的資訊檢索方式仍存在著許多問題需待克服。首先以目前的搜擎引擎來說，尚無法完全對網路上所有的網頁內容進行搜尋，其次從現行檢索的技術而言，仍以關鍵字詞檢索方式居多，並未涉及實體內容意義上的搜尋，在此基準下以資訊檢索中常用的「精確率」及「召回率」來進行檢索效能評估並不能真實反應實質內容檢索的成效。若要真實反應檢索的成效，應以使用者的觀點，以內容與內容間的詞義關係，來建立或還原資訊間相互的關連性。而主題地圖的研究就是希望讓資訊能有效的組織與鏈結，讓使用者能藉以瀏覽相關的資訊。所以主題地圖不單純只是用以做為瀏覽的工具，更能提供顯示或者探索新的知識之用，其所強調的是網路中各種數位內容相互間的關連性，並透過圖像的方式將其展現出來。

本論文以知識地圖為基礎的個人化知識管理、組織與瀏覽的架構，也實際引用數位典藏計畫中的 Metadata 來建置雛型系統，以驗證系統的可能性，希望能提

出一個人化知識組識與瀏覽的新架構。

關鍵字：知識表徵、個人化、詮釋資料、主題地圖

Abstract

Due to developer of electronic resource search technology, made more and more source to digitize, there are huge digital sources on the Internet. Users can use the search engine to query data they need. Because of the Information change and grown up, facing to such huge internet virtual database, human beings become to difficult get resource that they want. To search web pages or source in rustically way, can not address user's information need. Base on such huge and rich data, if system can provider a personal information search and tool, user can get information easy and quickly, will help him to made more success chance.

By using the search engine provider on the web, we can easy to find information we want, and use hyperlink to view the data, but it still has a lot of problem. At first, the search engine can not search all of the web pages on to internet. The second, the information research technology often provider a key word match, but not content march, so we can not use precision and recall score to evaluation query result set. We must in terms of user to build up the relationship between information bases on content. A topic map is use to organization and link the information, let user can browse the relation data. Topic map is not only a browsing tool, but also provider a display or explorer new information. It can render the relation ship between digital content and display it on graphic way.

In this paper, we want to build up a prototype system bases on the

knowledge map and provider personal knowledge manager, o organization and browsing architecture. We will use the metadata provider by national digital archive program to evaluation system and hope to provide a new architecture for personal knowledge organization and browsing.
Keywords:Ontology, Personal, Metadata, Topic Map

致　　謝

會開始想唸書，也是起因大學時，地科系的傅學海老師曾提到不要只是接受老師的講法，應該去挑戰及思考。感謝當年昭珍老師帶我進入了數位博物館這個大觀園，幫忙撰寫推薦函，讓我投考圖資系；而顯強學長更是一步步的教導我，從最早期的 Metalogy 系統的發展，最後還身兼論文的口試委員，學長那嚴謹且有系統思考和認真的態度，是我學習的好榜樣，當然，生存遊戲時就列為必擊中目標！

論文發展初期謝謝 LATTE 實驗室成員的支持與鼓勵，我還開了不少小差，把工作都推給了他們，而研究團隊本身的每週報告者，更是帶給我不少的點子。炳魁那些有趣的理論真是可以列為課外研讀方向，學勇三不五時的冷笑話提供調劑，而建華學長更針對題目本身提供了不少後續可以努力的方向，好友宇薇的支持與打氣，系上助教靜宜、小童和隆基的心情分享與戰術模擬。還有瑞瑩在旁的陪伴與鼓勵。

也因為有週遭老師與朋友的支持，才可以完成。最後，感謝我的父母，多年來的教導與哺育，才有今天的我。

順宏　2006.03.01

目錄

圖片目錄

表格目錄

第一章　緒論

第一節研究起源

網路中存在著上百萬甚至更多的網頁文件與資料,每天有超過三百萬人次的使用率，除此之外尚有更多的內部文件與資料存在於企業或組織內部，而這些文件與資料每天仍然以極快的速度持續成長與不斷的更新。面對如此龐大的文件與資料，人們也越來越難準確的搜尋到所需的資訊，更遑論能有效的去組織整理進而從中擷取所需的知識內容，讓使用者能從其中擷取所需的知識內容。

面對瞬息萬變的時代潮流，若能快速有效的掌握所需資訊，則可為企業或個人決策形成強而有力的支援。由於資訊科技快速進步，網路中多提供了搜尋引擎工具，讓人們獲取的資料相對變的容易與龐雜(但卻不一定能很準確的搜尋出使用者真正所需的資訊)。相對而言，組織整理這些龐雜的資料內容並能抽取出所需的資訊，並加以組織整合形成知識，變的越來越困難且耗時。現今網際網路的搜尋引擎(Search Engine)所使用的資訊檢索技術，仍侷限在就使用者所輸入的關鍵字詞進行檢索，比對預先儲存的網頁索引，並將所得結果排序，供使用者瀏覽點選。忽略相同詞彙在不同語境有不同用法，所以往往檢索出來的結果涵蓋絕大多數無用的資訊，對使用者知識的形成反而造成莫大的干擾。就知識的面向來看，知識的產生是藉由有用的資訊所整理萃取而成，倚靠的人的經驗及本身的成長背景，單純的文字比對所得之搜尋結果，並無法廣泛的涵蓋現今所有的知識內容。

知識管理強調對企業或組織中的知識做最有效的組織與管理,提供企業內員工彼此間知識的分享，並透過針對現有的知識內容予以統整分析，相互討論創造新的知識技能，以期能創造最大的商業價值。就知識操作的層面來說，單純的分

類與過濾，並無法顯示完整的知識架構，特別在於現今資訊爆炸的年代，大量的資料充斥於週遭環境，如何快速的篩選並找出自己所需的內容，且要能提供整體性的知識概觀架構和相關資訊脈絡，讓使用者能確實完整的檢索出所相關資訊，顯得更加重要。傳統單就文件本身的特徵(例如關鍵詞彙)進行分類及檢索，所查詢到的結果也只限於字詞的比對，無法真正是依使用者實際詞意的概念所欲檢索的相關內容。

知識之間彼此存在著一定的關連屬性，絕非單純且獨立的個體，隨著整體社會資訊化腳步的加快，從早期資訊難以取得，到目前可說是資訊氾濫的時代，每天在 WWW(World Wide Web)上不斷的有大量的資訊內容產生，並且人們可以很輕易的擷取使用，但目前的搜擎引擎亦無法取得所有 WWW 內上的網頁內容。除了從傳統的資訊檢索的「精確率」及「召回率」來看外，亦應處理內容與內容之間的關連，並嘗試建立或恢復其原本的關連。

第二節研究動機與目的

隨著邁入資訊化社會，知識具有更甚以往的高度價值，更是企業、組織強化競爭力和永續發展的要件。面對資訊時代所帶來的衝擊，企業或個人如何在最短時間獲取所需資訊，並藉以做為決策的依據，進而將知識分享給其他人，提升知識加值的效能，有效的組織與應用知識，將是企業或個人邁向成功之路重要的基石。現行多年的知識管理架構，大多針對內部文件及所產出的知識進行組織與管理，進而在企業內進行知識再造工程，以提升企業競爭力，更可減少因人員流動而產生組織知識的重大損失。在此過程中「個人」(或知識工作者)扮演關鍵性角色，藉由其類化與整理，將資料轉化為資訊，進而提升為知識，再將其回饋分享給其他人，形成一個知識加值回饋的迴圈。所以如何讓個人快速的擷取所需的資訊內容，是目前相當重要的課題。

現行知識管理系統，著重於傳統文件的分類與保存，並僅提供單純的字詞檢索功能，並無法針對個別使用者之所需來進行設計。亦即無法協助使用者針對資訊事件發生前後的關連性進行全盤性的了解，若能有效與有系統地透過資訊過濾與分類，並經由使用者個人的條件建置，將既有資訊內容，依條件內容做為分類，使其有層級與類別，將能協助建立個人主題瀏覽，並加速精緻資訊的擴展與延伸。

相對的，如何能在正確的時間地點內，提供個人其所需的內容，並協助其能夠有概括性的瞭解與掌握，則圖像的表示法可達到事半功倍。將知識的相關脈絡依照特定的主題圖像方式呈現，形成主題地圖，讓使用者可在最短的時間內，對該主題能有一全盤的瞭解，更提供其決策所需，並以建立一個良好的瀏覽介面來實現個人化。而根據文件組織架構的不同，使用者瀏覽的方式有下列三種模式：

一、扁平式瀏覽(flat browsing)

在此檢索結果文件集中，文件本身的分布散佈，而使用者在此情況下，逐一的去尋找文件及其相鄰文件或關鍵字，而部份使用者所感興趣的字詞，亦可以做為加入原有檢索條件，以重新組織並縮小檢索結果資料集。在這個情況下，使用者可以直接到找文件本身，但卻無法得到文件所屬結構的相關資訊。例如：搜尋引擎的查詢結構，可以讓人直接點選到該鏈結頁面，但卻無法告知該頁面所屬的父層結構(或原有鏈結名稱)。

二、結構化瀏覽(structure guided browsing)

在結構化瀏覽模式下，文件依循樹狀的方式來架構組織，實體文件位於樹狀結構中的末端位置(leaf node)。使用者依循樹狀路徑的

瀏覽方式，由樹的根部(root)開始在其相關臨近節點(node)間移動，並整體組織架構可知，文件本身的從屬關係。

三、鏈結式瀏覽(the hypertext model)

透過文件超鏈結的格式，使用者可以打破傳統文件循序閱讀的方式，在瀏覽資訊的過程中，可以輕易的依自己所感興趣的主題，「跳躍」式的來閱讀相關的知識內容，而文件本身亦保有其原有的結構與組織，藉此使用者可以更有效的尋找自己所想要的內容；但不好的鏈結設計，也可能會造成使用者在瀏覽的過程中迷失，而造成無效的資訊瀏覽。

所以良好的瀏覽方式設計，可以讓使用者快速有效的找尋到所需的資訊，透過完善的文件連結組織架構，更可以讓使用者在瀏覽資訊的過程中，一併獲取與搜尋主題相關的文件內容。

具體而言，本論文研究主要目的包含：

一、討論並解析知識主題地圖的應用及其目前所面臨的問題。

二、探討個人化的分類架構及使用模型建立之內涵。

三、以 Metadata 為例，由使用者的觀點切入 Metadata 的瀏覽模型發展與建置。

第三節研究範圍與限制

論文中所使用的資料其來源為國內數位典藏計畫所產出之 Metadata，藉由這

類已標示 XML(Extensible Markup Language，簡稱 XML)標籤語言的資料文件，來進行相關的研究與探討。因為相關研究涵蓋範疇相當龐大，我們將論文研究範圍限定如下：

一、研究中所使用資料之 Metadata，均已利用 XML 標籤語言進行描述，並不對未標記 XML 標籤的資料進行探討。

二、本研究重點在於建立個人化的知識瀏覽模型，並不對描述資料之語意標註方式(例如 XML 標籤)探討其合理性及適用與否，亦即不對標記規則的適切性進行討論。

三、文中僅針對欄位資料的內容進行探討，但不涉及實體資料欄位定義與結構，亦即希望所建構之模型，可通用於各種 Metadata 之資料結構。

四、本研究所探討之內容，以使用者瀏覽模型之建立為主，並不探討資料檢索方式及檢索策略之研究。

五、在資料庫建置方面，實驗中採 Microsoft SQL server 2000 作為測試平台，並利用 SQL(Structured Query Language, 簡稱 SQL)指令來模擬使用者瀏覽及檢索所需資訊。

研究限制如下：

一、研究中所採用資料來源，為數位典藏計畫中，典藏品之描述資料為主，資料本身已標記成為 XML 文件，所以在此並不涉及資料標記相關討論。

二、本研究所探討的方向，以模型建置為主，並不涉及系統平台的建立與開發實作。

三、本研究擬對資料的處理方式及使用者瀏覽方式作為探討，但不對資料如何運用作為討論。

第四節名詞解釋

主題地圖(Topic Map)

主題地圖是利用主題索引的概念，透過圖像階層(網路)聯繫的顯示，將主題、關連性及資源實體三者串聯起來並結合網路相互鏈結的特性，來達到讓使用者很輕易快速的瀏覽相關訊息。透過主題地圖能為全球資訊網路、區域網路、甚至是紙本所提供的資訊，創造虛擬的知識地圖。主題地圖實現長久以來各界欲創造使用單一格式表單來達成控制資訊的格式及編排的夢想，主題地圖即是利用格式表單控制資訊獲得及瀏覽的觀念，詳細敘述各種瀏覽層級，並模擬複雜的知識管理關係，提供個人化的資訊檢索的路徑（例如圖書館資源路徑），以幫助使用者更有效率的瀏覽電子資源。

知識主題地圖

知識組織關心的是如何讓組織的知識能夠有效地儲存以方便檢索，並將知識創造、分享給社群中的每一個使用者。知識管理（特別是在知識密集的產業裡）的目標有點類似將知識組織最佳化至某一程度，因為知識管理能夠確保組織內所有重要的知識資產和流程，能被人們了解分享並加以利用，以期對企業的營運上創造更高的價值與貢獻。而主題地圖扮演另一種知識組織的工具或手段，提供更友善的資訊瀏覽、檢索的服務介面。[1]知識主題地圖將相關或類似的主題，依照一定的型式呈現出特定的型式，充分表現出其脈絡和相關性，供使用者能有系統和有效率的獲取知識。

個人化瀏覽

使用者面對網路上蘊藏著數以億計的數位資訊，如何幫助使用者能夠快速檢索並瀏覽所需的相關資訊，對於知識爆炸的 WWW 是一個相當重要的課題。藉由與使用者之間的互動，記錄一些必要的訊息，以方便進行資訊的分類，進而提供符合以個人需求為導向的服務。以網際網路的發展為例，網際網路資訊服務提供者，不論在於強調資料的搜尋的廣度的搜尋引擎如 Yahoo[2]；或者是著重在於專業知識的整合與深度如 CNN[3]及 eBizPort[4]，均開始提供個人化的服務，如 Yahoo 的 My Yahoo![5]，Yahoo 更甚為該地區建立區域性資料[6]，提供搜尋當地相關資訊。並藉由此一服務的提供，提高使用者的忠誠度[7]，更便利使用者資訊搜尋。

Metadata

Metadata 一般泛稱為描述資料的資料[8]，主要是描述資料屬性的資訊，用來支持如指示儲存位置、資源尋找、文件紀錄、評價、過濾等的功能。此一定義係源自在一九九五年三月由 OCLC（Online Computer Library Center）、NCSA（National Center for Supercomputing Applications）兩單位共同主辦名為「Metadata Workshop」研討會，廣邀圖書館學、電腦科學、文獻編碼、以及相關領域學者專家等參加。在此會議中，首先提出了「資料的資料」作為 Metadata 的定義，自此之後，有關 Metadata 的各種定義亦紛紛出現例如圖書館自動化系統所使用的機讀格式(Machine readable cataloging，簡稱 MARC)，即為一種 metadata。國內對於「Metadata」現有的翻譯名詞有元資料[9]、超資料[10]、詮釋資料[11]及後設資料等。而在本研究中對 Metadata 的使用及定義為「資料輸入的格式規範」。

XML

XML 為 1998 年 2 月由 W3C[12]所公布的一種標籤語言(Markup Language)，為 SGML(Standard Generalized Markup Language，簡稱 SGML)所精簡而來，具有簡單、可擴充及高度結構化的特性。

XML 為 SGML 精簡後之子集合，並非如 HTML 由固定的標籤(Tag)格式所構成，其允許使用者自行定義標籤來表現各種不同意義的標示，並可應用於各式領域。XML 標籤的定義和語言及其操作平台無關，並能提供定義於 Web 上的結構化資料交換格式。而透過對於標籤的操作，讓應用程式可以很容易將 XML 文件中的資料分離出來，並進行利用。

基於以上的特性，自動化系統都能輕易的修改符合處理 XML 文件功能需求。在實際應用的例子中，XML 已應用於諸多電腦應用上(如 SOAP 及 OAI)，不僅可以提供整合不同類型的文件，更可以作為機器之間交換的標準格式。

[1]Alexander Sigel M.A.(n.d.),Towards knowledge organization with Topic Maps(http://index.bonn.iz-soz.de/~sigel/),pp1.
[2] YAHOO(http://www.yahoo.com)
[3] CNN(http://www.cnn.com)
[4] eBizport(http://econage.com/product/eworkport.htm)
[5] My Yahoo! (http://my.yahoo.com/)
[6] Yahoo local(http://local.yahoo.com/results)
[7] 施毓琦，「大學圖書館網站個人化服務之使用者需求研究」(國立臺灣大學圖書資訊學研究所，碩士論文，民 92 年 6 月)，頁 1-2。
[8] Stuart Weibel, Jean Godby, and Eric Miller,OCLC/NCSA Metadata Workshop Report(http://www.oclc.org/oclc/research/conferences/Metadata/dublin_core_report.html)
[9] 吳政叡，「從電子檔案和元資料看未來資料著錄的發展趨勢」，中國圖書館學會編，海峽兩岸圖書館事業研討會論文集(台北市：編者，民 86 年)，頁 163-174。
[10] 陳昭珍，「電子圖書館資訊組織問題之探討」，中國圖書館學會編，海峽兩岸圖書館事業研討會論文集(台北市：編者，民 86 年)，頁 175-196。
[11] 陳雪華，圖書館與網路資源(台北市：文華，民 85 年)，頁 206。
[12] World Wide Web Consortium, W3C (http://w3c.org/TR/2004/REC-xml-20040204/)

第二章文獻探討

知識管理可以讓知識能被有效的運用，藉由主題地圖工具對知識進行組織與管理，提供使用者彼此知識交流的工具。知識組織關心的是如何將知識儲存、整合與並能提供使用者精確快捷的檢索方式，使成員間能達到知識共享。透過主題地圖，可以藉由使用者自已來定義組織所需的知識結構，並將之轉變成為可以相互分享的 Metadata 資訊。而目前多以 XML 方式來作為定義主題地圖的標記語言，它能有效的提供描述知識組織的新的方法，在此相同規範標記語言下使用者能用更少的時間減少不同 Metadata 之間創造(creative)、維持(maintenance)與交換(interchange)。又主題地圖可以讓使用者依個人不同的知識需求觀點提供彈性的組織結構，亦即其在設計上特別著重於使用者導向[1]，透過知識地圖的建立我們能對知識予以個人化的組織分類，對於使用者在資訊的擷取與組織有相當大的助益。

本文擬以數位典藏內容網站為例，探討如何針對不同使用者提供個人化內容瀏覽服務，建立使用者本身之知識地圖。在文獻的分析部份將就：一、深入了解主題地圖及其發展現況，並探討內容型式及其應用層面；二、針對知識管理與主題地圖在實務應用上結合的討論；三、從資訊科技的角度，探討自動分類應用於 Metadata 資料的情況。

第一節知識需求的發展

「知識即是力量」說明了能掌握運用資訊的人，具有較強的優勢與競爭力。資訊科技的發展加速了知識的傳播與創造，網際網路已成為人們生活中不可或缺的要素，然而面對網路中蘊涵的龐大資訊，如何有效且快速的組織與管理，進而產生有用的知識，將是使用者所面臨的挑戰。而知識組織可以將複雜的概念有系

統的加以組織，讓大家便於取用。知識組識是人們將資訊價值加入知識的集合物中，屬一種跨學科領域活動。[2]知識管理所面臨的最大挑戰即是隱性知識的分享。組織溝通結構的良好與否，對於知識管理在組織的發展，具有決定性的影響。面對大量成長的資訊內容，應善用資訊科技處理、過濾及分類資訊，縮短處理時間並提高效率以及促進交流。

知識管理領域中以知識可現程度來區分，主要可分為內隱知識及外顯知識兩大類，而外顯知識即為表現在外的知識，可以文字、數字、圖形或其他符號來定義屬之，如書本知識、手冊…等；內隱知識多為高度個人化，通常只可意會不能言傳的知識屬之。如：直覺、價值觀與基本假設，以及技巧及個人專業知能、經驗…等，存在於個人標準內，而難以規範或直接取得。Nonaka & Takeuchi[3]認為知識的創造是經由內隱(Tacit)與外顯(Explicit)知識互動而得，有社會化、組合、外部化與內部化四種不同的知識轉換模式。

知識的分類目的是將無次序或無組織的文件內容，依一定的原則及方法，依個別文件的特徵進行類化與歸納，以方便提供使用及傳播。知識的取得與創造，需經由一連串處理與內化的過程，透過對知識的組織分類，可以加速提供使用者轉換知識為已所用。在知識的取得方面 Gilbert & Cordey-Hayes(1996)[4]曾提出一個知識管理五階段架構，包括：知識獲得、溝通、應用、接受及吸收。也就是組織可以由外部獲取所需知識或由組織內部自行創造而得。對於所獲得的知識，組織形成溝通的機制與分享。進而能將所獲取的知識加以應用知，促使組織成為學習型組織，並讓所有組織的成員都能接受。知識的移轉與分享，最後能將知識內化，產生更大的效能。

科技的發展讓資訊的取得方式更加多元化，而資訊取得後需經由個人的處理轉化為知識，而隨著每個人的使用情境不同而會有不同的處理結果，所以知識並無法單純的直接轉移給任何一個人，而是需要一些轉化的工作。

與一般實體或具像的事物相比，知識或資訊相對而言是虛無、抽象的概念，對於知識的描述有時很難以用文字或言語加以敘述清楚，而其內容更是因人而異，每個人希望了解的知識又多所不同。如果知識是已經驗證過的資訊，那麼資訊就是將資料詮釋成一有意義的架構。[5]資料是原始的數字及事實，資訊是處理過的資料，而知識則是資訊的行動化。知識管理的目的在於將有用的資訊做適切的組織、管理，以期能產生最大的收益。所以如何提高知識的蒐集、儲存、分享進而創新，產生更具競爭價值的知識或智能，並減少過程中所產生的知識的遺失或損毀，是知識管理所著重的課題。透過知識管理可以有效的利用現有資源，並發揮其存在的潛在價值，提供知識的再利用與知識的探索，創新更具價值得新知識，以提供工作決策與流程的支援。

對於知識而言，知識本身又具有特定性。知識的產生是經由對最初資料的收集與整理而成的資訊，再經由使用者的個人詮釋、思考、歸納等而產生。知識本身已經由個人的彙整粹取資訊而成，對自己本身而言是知識，但對另一個使用者而言，可能只是單純的資料。概因所面臨要解決的問題也不盡相同，每人的起點行為亦不同，所以知識本身是相當個人化且具抽象性質。[6]

Churchman[7]對於知識的概念化與對「知識存在於使用者而不是資訊的集合」之陳述相一致。其認為從個人化的角度來觀察，如果要使個人的知識對於另一個體是有用的，它必須是可以一種交換的方式，將知識傳遞給對方，以便使對方得以理解與獲得。且儲存的資訊並無法產生任何的價值。唯有個人經由思考進行組織整理、運用、分享、創新、學習、反饋的過程之後，資訊才能變成有用，因此，知識管理是指有系統且有組織地獲取、組織和傳遞正確資訊之特定過程，使人們可以更有效的從事工作之生產。

知識管理是多面向的，也就是說，並不只有科技才是知識管理系統的核心，其他像是文化或組織的議題也涵蓋其中。事實上，文化和知識議題的有效解決方

法也是知識管理系統所關注的發展方向。[8]文化是有生命的，那就是之前數位化所強調的議題。事實上，文化和知識議題的有效解決方法也是知識組織在發展時所關注的焦點。因為人們往往只以自己是否使用科技來做為行為理性的判斷標準，卻忽略了目的的內涵往往是要藉由多面向的參照也因此經常顧此失彼。

由知識管理系統面來看，現存知識管理系統多以企業組織內部的知識組織為導向，而較少涉及個人層面之需求，鮮少論及知識彼此間的相關脈絡關連，但對知識的產生而言，最重要事即在知識脈絡關係的建立上，而非單就片斷的內容即可供人使用，現行知識管理系統在資訊的處理上仍有一些不足[9]。

現行系統多以關鍵字導向的檢索方式實作，所憑藉是關鍵字詞表面來進行比對而非字義。所以該字詞若不存在於資料集內容，或資料集是採用不同的詞語來表示相同的概念的詞義，系統便無法順利擷取到相關所需資料內容。所以在傳統的檢索架構下，資料內容缺乏字義關聯訊息，純就字形進行比對檢索，將可能會有高召回率，但精確率卻相當低，因為在此架構下使用者可能檢索到大量其不感興趣且無關的資訊內容。人們是依靠瀏覽及閱讀來獲得有系統的知識內容，再將所獲得的資訊加以組織創新，產生新的知識，但就目前檢索的方法，使用者並無法有效的將分散於各地的資訊透過簡單、個人化的方式，進行檢索瀏覽，更遑論能加以組織，創造新的知識供人使用。

當數位化的腳步加快，使用者會發現周遭的資訊以爆炸性方式成長，對傳統的知識管理系統架構而言，也變的更形複雜且難以維護，相對而言，要讓內容資料保持正確而一致性且做經常性的更新，變的更加困難。另外目前知識管理系統多缺乏個人化的概念，無法針對單一使用者的需求，提出其所需的內容及其相關部份。僅能單就系統設定的方式，針對各使用者提供相同的內容供其使用，同時，內容本身亦無法直接經由機器再處理使用。

網路的發展，帶動知識發展與成長更加多元化，除了傳統以文字方式呈現的內容，更有以多媒體視訊等方式呈現的內容，而相對來說結構本身亦更加的鬆散且整合多種數位格式檔案，而使用者針對相同的內容，亦會有因其目的的不同而有著不同的需求，除了考量於內容本身的處理外，更應要考慮使用者個別化的影響。

第二節主題地圖的應用

主題地圖是用來描述知識的一種方式，將相關的概念群聚一起，建立起相互的關聯性。並如同書後索引提供了使用者能快速檢索相關資料的方法，書後索引將某一主題及相關附屬主題，依序建立關連索引，同時也建立「見」(see)與「參見」(see also)的關係讓所列的關鍵詞(主題)間建立起相互關連性。而主題地圖本身更利用了電腦工具所提供強而有力的資料瀏覽能力，除了可以透過圖像表示法讓人了解各主題之間的相互關係之外，更可以直接透過 XML 標籤語言，直接連結相關資源所在，在瀏覽的過程中不會脫離原本的主題架構，可說是將主題與資源緊密結合。而主題地圖本身帶有豐富的語意訊息，可以有效的用來組織知識架構，並且表徵、推理與解決相關大量無序資料所帶來的問題。

主題地圖的目的是希望能夠提供符合使用者資訊需求的資源整理及組織方式，跳脫傳統平面式、二維式的資訊組織模型，回歸多層次主題關聯架構，讓使用者可針對某一主題及其相關資源進行整合性的瀏覽。透過主題地圖可以讓使用者依其個人化的資訊需求行為來瀏覽特定概念的內容，而不會改變既有資料本身的內容或結構。這如同圖書館中的書目卡片與書本之間的關係，書目卡片的排列方式，可以依照人名、書名或主題而有不同的排列方式，但書本的實體位置卻是固定的，使用者可以依照自己所需的卡片排列方式來檢索書目，來查詢書本的放置架位。相同的，主題地圖與資源間的關係，就好像書目卡片

與書本，使用者可依本身對資源瀏覽需求的不同，透過針對同一資源不同定義的主題地圖，將資源以不同的面貌來呈現。透過主題地圖的協助，人們可以輕易的將資源進行適當的組織，以期能讓使用者快速檢索到相關的資源。

主題地圖的架構主要可分為三個部份 Topics(主題)、Associations(關聯)、Occurrences(資源指引)；主題(topics)是源自於希臘字 topos，代表的是 location 和 subject，主題可以代表任何名詞，是主題地圖中的最小類別歸屬單元。使用者可以透過事先定義與其應用相關的 topic 來進行使用，透過類別標示及名稱來「定位」其位置，並經由範圍限制其應用的領域。Associations(關聯)是描述二 Topic 之間的依存關係，而且是唯一，association 可以用來含括一或多個 topic。再針對各主題之屬性及特色找出其中的關聯處建立其相關性。Occurrences(資源指引)是資源與 topic 之間的關係。而資源可以是電子文件、實體圖畫，或任何資料，且資源本身可以是包含於 Topic maps 內部或外部資源。如圖 2- 1 所示。

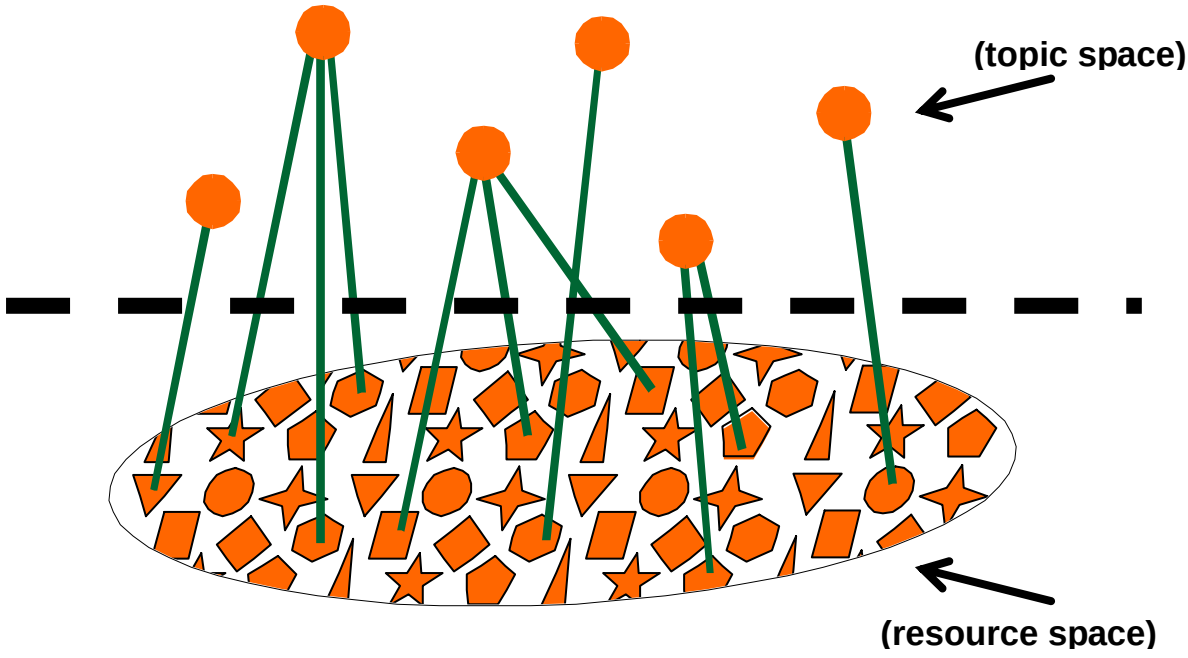

圖 2- 1：主題與資源之間關係例圖。(Bishop, D. 2001)

同樣的，主題的類別、關聯和資源指引亦是一種主題。

而主題地圖可以應用在主題的標示與識別、出版文件的標示和整合各式各樣的知識[10]。藉由指定主題的應用範圍，可以針對單一知識本體，而有不同的觀點或處理方式，並可以追蹤知識整合的過程。[11]主題的應用範圍讓單一的主題地圖提供多種的觀點及提供主題地圖本身的應用範圍限制，以免造成濫觴，並追蹤知識的整合與個人化的知識處理，能提供依使用者本身的能力、需求或對知識的存取安全等級，給予不同的結果。

藉由主題地圖本身所帶有的語意及其了解其所應用的範圍，則主題地圖本身對使用者而言能以人的思考方式，建立起主題與資源之間的關連，並提供瀏覽的協助；至於對應用程式的輔助上能對應用程式提供語意的查詢，相對於傳統的查詢，係對欄位的內容(context)進行字串本身的比對，亦即使用者所鍵入的關鍵詞需存在於資料庫內容，才能找到資料。對於不同需求的使用者，或現有資料，亦可以提供不同的面向與服務，並整合現有內容，而不需使用者的介入。

所以主題地圖本身不單只是描述主題本身的相關關係，更可呈現資訊整合與交換的一種方式。透過主題地圖本身，不但可以有層次且依序的瀏覽資源內容。

第三節自動分類

文件分類的目的，在對文件進行分門別類的加值處理，使得文件易於管理、利用。文件分類可將非結構化的資料(文件)，透過存在於文件本身的特徵，進行人工或自動化的分類，將意義相近的文件叢集化，是資訊組織、主題分析與知識管理的重要工具。[12]同樣，知識的吸收，也透過整合與分析資訊後產生。如何有效的加速資訊的處理，同樣的也就能減少處理的時間。近年來，數位文件不斷的大量累積，數量已大到無法像傳統一樣，採用人工進行處理與分類，或者說，用

人工處理反而顯的不具有時效性與經濟性，自動分類處理也因此應之而生，特別是近年來在自動分類的成果上有著顯著的發展，[13] [14]分類演算法的發展亦提供良好的階層式分類架構[15]，相對的也提供分類一致性與架構上的高擴充性。

文件分類，先要瞭解文件的主題大意，才能給定類別，因此是相當高階的知識處理工作。要將文件分類自動化，必須先整理出分類時的規則(或特徵)，電腦才能透過所定義的特徵進行自動分類(叢集化)。然而，因為文件屬於非結構性資料，其分類規則(或特徵)通常較難以人工分析整理的方式獲得。因此，透過機器來做自動分類是從事文件分類研究努力探討的課題。自動分類是否成功的一個重要因素，在於是否能從訓練文件中學習(或抽取)文件分類所需的特徵向量，再透過從訓練語料中抽離的特徵值，對新的文件特徵值進行叢集相似度計算。就目前自動分類的最常使用的特徵值為關鍵詞(keyword)，而擷取關鍵詞的做法，可由統計模型來計算詞彙頻率及相互關聯性來建立分類文件的關鍵詞集(keyword set)。

自動化的擷取關鍵詞可以節省大量的人力與時間，並可以短時間內獲得大量的關鍵詞，但所取得的關鍵詞可能較不精確，從文獻中大致上可分為詞庫比對、文法擷取與統計方式。

詞庫比對法是藉由現有已知的詞庫，透過對於文件本身進行斷詞處理作業，並擷取其中的特徵詞彙，做為後續使用。這個方法，需先建構一個關鍵詞庫，方可對文件內容進行詞彙比對。早期建置詞庫需要大量耗費的人力與物力，而所建構出來的詞庫，並無法完整蒐錄相關語彙，現在雖可透過自然語言處理技術，快速且大量的從語料中擷取相關關鍵詞彙，但又常因為所採用語料的關係，對於不同相關領域的文件，效果即大打折扣。同樣的，詞庫比對法，是採用存在於詞庫內的詞目，對於文件進行比對，對於已知詞的擷取是相當有效，但面對現今數位文件的大量生成，未知詞的數量更是不勝枚舉，且隨著新詞的逐漸增多，詞庫越

來越大，比對的速度越慢。亦即詞庫比對法，在未知詞與專有名詞的處理上並無法做有效的擷取或判斷。

文法擷取是利用人類語言本身的特性，如文法及構詞規則，對文件進行剖析並進行語法標記，並利用一些方法及訓練的準則，過濾不適合的詞彙。[16]但此種方式，仍需事先針對語言特性建立文法或構詞規則，或者是採用內含標記之語料庫，所以其缺點亦和詞庫比對法類似。此外，部份程式甚至只能解析符合文法之句子，對於短標題或書目資料，則無法正確運作。

統計方法是建立在文章中的重要的關鍵詞句，經常會重複的出現或被反覆引用。[17]利用簡單統計量數來計算文件中詞彙出現頻率及詞彙共現(co-occurrences)關聯等性質，找出文件中重要的關鍵詞彙。統計方法最大的優點在於不需由專家事先建立詞庫或文法，藉由統計量數與門檻值的過濾，即可有效抽取蘊含在文件中的重要的關鍵詞彙集，而統計方法與文件本身是用何種語言書寫並無關聯性，所以可以廣泛應用在各國語言文件的處理上，這種方法在自然語言處理中對於新詞或未知詞的發現有相當不錯的成效，但由於其效果受統計量數與門檻值等參數的影響，所以結果可能會包含一些雜訊內容，但只要透過嚴謹的統計模型設計與控制，將可大幅降低雜訊的影響。

除了以上幾種方法外，另外有結合文件本身特性，諸如：關鍵詞位置、排版原則(粗體、括號或破折號)及文件內含的標記，來進行關鍵詞之判斷擷取，或是結合數種方式進行。另外，輔仁大學曾元顯老師(2000)亦提出，利用最大重複數的原理─最大關鍵詞法並結合統計方法與「位置串列」的優點，進行關鍵詞的擷取，概述如下：

最大關鍵詞法[18]是一套能擷取所有相異的最大可重疊重複片段技術。在此的「最大」指的是在文件中重複字串長度最長，或出現頻率最高的意思，亦即，某

一個重複片段若不為任一重複片段之子字串時，就應該要被擷取出來。舉例來說，以表 2- 1 文字為例，則擷取關鍵詞彙步驟如下：

表 2- 1：範例文字

球迷熱賭盤瘋 全台賭金數十億

【記者蔡政諺／高雄報導】

中華職棒象牛年度總冠軍賽昨天在澄清湖球場開打，場內球迷熱情不減，場外賭盤更瘋狂，雖然傳出檢警搜索簽賭盤口，盤口還是開出「兄弟贏興農一分屬和局」的賭盤。林姓組頭表示，估計全台簽注金額達數十億元。

本月八日爆發中信鯨球員簽賭醜聞後，檢調隨即鎖定中南部的職棒簽賭組頭偵辦，中南部組頭連日來紛紛避風頭，緊急變更盤口聯絡方式，直到昨晚象牛年度爭霸戰開幕賽前一個半小時，才避開檢警的查緝，接受賭客下注。

高雄地區林姓組頭透露，昨晚牛象之爭雖然兩隊實力相當，但組頭普遍受到兄弟象球迷氣勢如虹的影響，開出「兄弟象贏興農牛一分屬和局」，即比賽結果若是兄弟象剛好贏興農牛一分，盤口退還賭客下注金，不計輸贏；但若兩隊打成平手或兄弟輸球，則賭兄弟隊的輸了；反之，若兄弟象贏興農牛兩分以上，則下注兄弟象的賭客贏錢。

林姓組頭說，雖然日前傳出職棒打假球事件，但包括組頭、簽注者或球迷都認為，總冠軍賽關係球隊排名，球員為爭榮譽，加上牛象兩隊嚴格管理球員，不可能打假球，賭風因此更盛，下注一萬元組頭抽五百元，贏了組頭賠九千五百元，下注一、二十萬元者大有人在，全台賭金估計達數十億元。

【記者李一中／高雄報導】中信鯨隊有球員疑似涉賭，使得中華職棒總冠軍賽的執法公正性格外受到關注。中華職棒裁判組長周濃舜昨天表示，裁判在總冠軍賽期間將嚴格限制個人行動，「即使去上洗手間，也要跟我報備，違規者將調查嚴辦。」

1. 統計相鄰二字串(2-gram)之頻率，可得如下結果(節錄)

(...上牛:1,牛象:2,象兩:1,兩隊:3,隊嚴:1,嚴格:2,格管:1,管理:1,理球:1,球員:3,不可:1,可能:1,能打:1,打假:2,假球:2,賭風:1...)

2. 設定臨界值為 1，並去除低於臨界值之字串，可得：

(…x,牛象:2,x,兩隊:3,x,嚴格:2,x,x,x,球員:3,x,x,x,打假:2,假球:2,x...)

3. 考慮字串所在位置後，組合相鄰二字串即可得：

(…x,牛象:2,x,兩隊:3,x,嚴格:2,x,x,x,球員:3,x,x,x,打假球:2,x…)

4.整合所得之詞彙即可得(牛象、兩隊、嚴格、球員、打假球)等等關鍵詞。

透過上述方式從文件擷取的關鍵詞，詞彙間並不具有關聯的資訊或上下位關係，雖可用來作為文件的特徵，可在檢索上加以運用，但卻無法明確定位文件本身在文件集中的位置。相對來說，人工編輯之索引典則是具有階層性及詞彙間相互間的關係，透過索引典的使用，可以定位某一關鍵詞在實體檔案關係中的「位置」，並據此將所屬文件亦歸至該類別或過濾掉無關的文件，檢索出實際相關的文件。

透過前述的自動擷取與自動化分類，可以在短時間內獲得分類階層，但自動化處理的方式著重在以資訊檢索的角度，是以「關鍵詞」為基礎，而非以「內容」為導向，所以分類的結果往往缺乏語意，而無法滿足使用者的需求。自動分類的方式，可以用在縮短人工分類的時間[19]，但現暫無法完全取代人工分類，若能將

人工分類所產生之訓練資料供分類器參考，則應可以獲致較好的成果。也因每個人對資訊的需求不同，亦導致相對欲探求的知識內容往往南轅北轍，所以如何透過個人化的處理讓使用者能快速取得所需的資訊，並能從其中擷取重製產出個人所要求的知識內容，是知識工作者努力的目標。由前述可得知，知識本身具有獨特性，而現今知識管理仍是群體為導向，較不涉及個人層面的需求。主題地圖的發展將可帶動賦予知識內容多元化面貌呈現的嶄新模式，如何能讓使用者透過主題地圖，針對知識本體進行個人化，加快知識的吸收的速度，或得到概括性的全盤掌握，則為本研究最主要的目的。

[1] Alexander Sigel M.A.(n.d.),Towards knowledge organization with Topic Maps(http://index.bonn.iz-soz.de/~sigel/),p4.

[2] Sigel, A. (2000a). The knowledge organization on Internet mini-FAQ(http://index.bonn.iz-soz.de/~sigel/ISKO/wiss-org.faq.html)

[3] Nonaka, Ikujiro and Takeuchi, Hirotaka, 1995. The Knowledge-Creating Company. Oxford University Press, 284 pp. ISBN 0-19-509269-4.

[4] Gilbert, M. and Cordey-Hayes, M.,"Understanding the Process of Knowledge Transfer to Achieve Successful Technological Innovation,"Technovation,16(6),pp.301-302, 1996.

[5] Vance, D. M.(1997), "Information, Knowledge and Wisdom: The Epistemic Hierarchy and Computer-Based Information System", Proceedings of the 1997 America's Conference on Information Systems, http://hsb.baylor.edu/ramsower/ais.ac.97/papers/vance.htm

[6] Maryam Alavi, Dorothy E. Leidner(1999). Knowledge management systems:issues, challenges, and benefits. Communications of AIS Volume 1, 1999 Article 7

[7] Churchman, C. W. (1972) The Design of Inquiring Systems: Basic Concepts of Systems and Organizations, New York, NY: Bencis Books.

[8] Maryam Alavi, Dorothy E. Leidner(1999). Knowledge management systems:issues, challenges, and benefits. Communications of AIS Volume 1, 1999 Article 7

[9] Dr. John Davies, Professor Dieter Fensel, Professor Frank van Harmelem(2004). Towards The Semantic Web,2004, John wiley & Sons, LTD. P3

[10] Pepper, S. (n.d.). Topic Maps: The Tutorial. Retrieved Aug. 11, 2004, from http://www.ifi.uio.no/in-id/lecture_slides/pepper_advanced_concepts.ppt

[11] 同註 10。

[12] 曾元顯。(2004)。文件內容自動分類。(http://www.lins.fju.edu.tw/~tseng/ResearchResults/categorization.htm)

[13] Jain, A.K. and Dubes, R.C.(1988) Algorithms for clustering data, Prentice Hall.

[14] Jain, A.K., Murty, M.N., Flynn, P.J.(1999) Data Clustering: A Review, ACM Computing Surveys, 31, 264-323.

[15] Daphne Koller, Mehran Sahami "Hierarchically classifying documents using very few words." Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997.

[16]陳光華,「資訊檢索查詢之自然語言處理」,中國圖書館學會會報 57 期(85 年 12 月),頁 141-153。

[17] Paolo Tonella, Filippo Ricca, Emanuele Pianta and Christian Girardi, Using Keyword Extraction for Web Site Clustering

[18]曾元顯,數位文件之資訊擷取與檢索。(台北市:全壘打文化事業有限公司,民 89 年 9 月),頁 269 頁。

[19] Chao-chen Chen, Jian-hua Yeh, Shun-hong Sie, "Government Ontology and Thesaurus Construction: A Taiwan Experience", in Proceedings of the 8th International Conference of Asian Digital Libraries (ICADL2005), Bankok, Tailand, December 2005.

第三章研究方法

第一節系統分析與資訊處理

在本研究所要實現的系統主要是針對來自數位典藏國家型科技計畫之資料，讓一般使用者可以透過簡易的分類設定，將原始資料進行組織、分析與瞭解，轉化為其所需之知識內容。透過系統所提供的個人化知識組織，將原始文件分類進行重新組織與分類，以符合個人使用之需求，並提供簡易之瀏覽方式，供其瀏覽且能綜觀整體知識內容。

數位典藏國家型科技計畫之原始資料非常的多樣化，各典藏子計畫所採用之描述標準亦相當的不一致，故系統在設計之初需對 Metadata 進行結構分析，依據其結構上的特性，提出一個與資料本質無關(Data Independent)的瀏覽結構。現行的 Metadata 格式種類繁多，以國內數位典藏相關計畫為例，即有數十種不同的格式存在，而各種 Metadata 的標記欄位又不盡相同，如何有效的利用既有的欄位，進行資訊的分類及瀏覽，更甚提供不同 Metadata 格式間的相互對應，縮小彼此間的差異門檻，而可互相相容操作，則為一個重要問題。因此系統的設計之初，應要考慮能相容操作多種格式之 Metadata，透過系統而能達到與資料本質無關之瀏覽操作。

隨著資訊呈現爆炸性的成長，其需求卻日益的增加。自動分類的目的，在於可以輔助人工分類的不足，大量的減少人力與時間，協助人們發掘知識，而知識是依個人因時、因地、因物之不同，而有不同的需求與面向。系統應能提供個人化瀏覽模式之模型，而如何才能滿足此一需求？經由系統開發流程方式，先分析使用者的需求及資料內容，實際繪製各式相關圖表，和規劃操作流程。系統分析階段的目的是為了解針對個人化知識主題地圖的需求，並且建新系統的邏輯模

型。依據所收集文獻資料及使用者需求分析，建立需求模型(requirements modeling)，並評估所需之相關技術及背景，以期能滿足現行需求，並可繼續提供未來發展使用。(如圖 3-1)

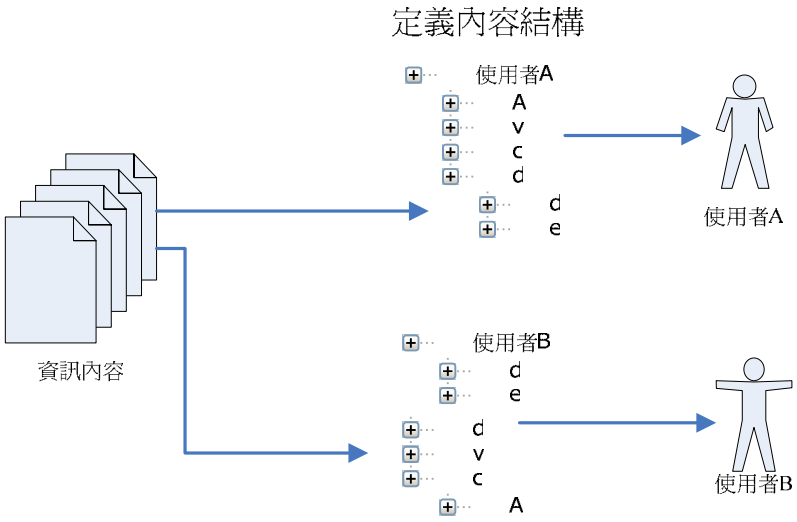

圖 3-1：系統概觀

第二節研究架構設計

本研究首先需界定了系統的需求面及實作範圍而進行實作雛型系統，驗證理論的可行性。透過雛型系統的開發與實作，實際分析及探討可能性，了解系統實際上線後，可能會遭遇到的問題，並一一進行調整，修正理論的內容。並依據開發過程所發現的問題進行改善，使雛型系統能更符實際所需。

透過文件自動分類的方式，輔助人工分類的進行，協助使用者建立個人化的瀏覽模型。透過使用者的操作，並可將已組織好的內容進行分享與引用他人所建立的架構。但由於使用者之個別需求不同，針對相同測試資料集，所進行的內容呈現方式亦會有所不同。本研究針對個別使用者分類之資料，進行自動分類成效評估，進而求取最佳分類參數值。

針對單一的資訊內容，本模型應能提供不同觀點的使用者，依各自所定義之結構而有不同的呈現結果，並能進行結構化的瀏覽方式。而單純的分類並無法建立起各類別之間的關連性，需要再建立類別與類別之間的關連，使用者方能對此建立整個內容的概觀。

第三節分類階層的建立

分類階層的建立是相當的耗費時間與人力，而在持續不斷的建立過程中，隨著時間的遞增，更難保持著一致性，同一個人在不同的時間點所進行的決定，亦可能會有不同的結果產生。對使用者來說，同時維護大量且巨大的分類架構將變的更加困難，但又因人所處理的資訊往往帶有強烈的語意或訊息，單純採用機器自動分類，往往會失去其應有的精準度。

人工建立個人的分類階層雖然有一定的限制，但透過一些自動分類的輔助，以人工分類的資料作為訓練資料集，對未分類之文件進行分類，應可以提高一定的精確度，特別在大量資料的採用上，透過人工建立少部份的分類資料作為訓練資料，來進行自動分類，並輔助以人工的確認與修正錯誤的分類的內容，不但可以減少時間，並可以確保分類的一致性。

針對本研究的需要，我們設計出一個可依個別使用者所建立的分類結果，以其作為訓練資料內容，並對未分類的資料進行自動分類，且提供使用者一確認的機制，作為修正分類結果所需。而又因個別使用者間可能存在著不同的分類方式，故系統亦採用一動態門檻值的方式，針對不同的使用者的分類結果，提供不同的門檻值，作為文件分類判斷的依據。

第四章 系統實作

第一節系統架構

在系統實作上，我們以 Microsoft SQL Server 作為資料儲存平台，透過 XML 標記語言作為資料相互交換的標準格式。在資料部份引用國家數位典藏計畫產出的 Metadata 詮釋資料作為實際分類操作內容，因典藏資料建置的格式標記並無統一規範標準，所以系統必須允許滙入各式資料定義格式(Metadata)，並能對其進行操作為主，唯資料本身需為 XML well-format 格式，且匯入的 Metadata 需包含資料的 DTD 或 Schema 的標記定義，並據以規範其所承載內容。系統透過滙入資料的 DTD 或 Schema 定義，建立 XML 結構綱要表，並作為 XML 資料驗證(verify)之用，以確認所載入之 XML 資料標記格式一致。

從系統設計的考量上，為達到透通性(transparent)與可移植性(portable)，選擇 JAVA 作為系統開發環境，並以 JDBC 作為 JAVA 程式與資料庫之間連結協定，透過 JDBC 提供對資料庫的操作，達到程式與資料庫獨立的功能。[1]

在考量其複雜性和降低困難度的前提下，我們依功能將系統分解成多個獨立的模組，在整體架構上使用共同的模組架構進行實作，可以動態的配置加入或移除功能，在操作上相當容易並提供高效率的客製化介面，以及後續的更新和維護作業。在資料架構上，系統把資料分為「資料儲存」、「資料格式定義」、「資料呈現」三個層面如圖 4- 1，其中實體資料儲存於 SQL server 內，並透過 JDBC 進行操作。

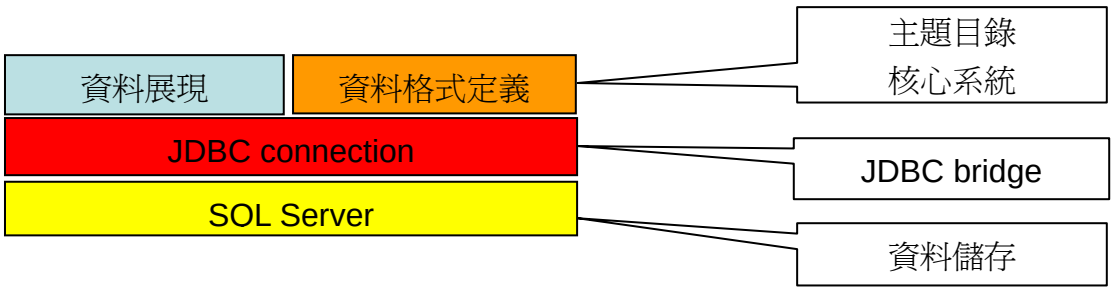

圖 4- 1：系統架構

針對使用者擬欲進行建立分類之 Metadata 來說，系統需能提供相容各種 Metadata 之資料定義架構模式。現今 Metadata 種類繁多，Metadata 之間格式及資料建置方式又不盡相同，且彼此之間的轉換易造成資訊的流失，在此系統只處理以 XML 標記處理後之 XML 文件，作為主要建置資料來源，並且，系統擬不對 Metadata 彼此之間的格式進行轉換，即不提供由一種 Metadata 轉換至另一種 Metadata 格式，故亦需建構能相容於各式 Metadata 的平台，並在其上提供各式操作服務。

而就 XML 文件本身屬於階層式架構，對於關連式資料庫中的表格連結方式來說，是相當不同的。所以要利用 SQL server(關連式資料庫)系統儲存 XML 文件，需在不影響到資料本身的情況下重新規範相對應的欄位定義，以避免資訊在轉換格式時的流失，並提供完善且便捷的操作能力。

	MRN	PARENT	PID	TAG	ID	DATA	ATTR_FG
1	10	*	-1	record_list	0		0
2	10	creator	11	arranger	14	蘇森墉	0
3	10	creator	11	composer	12	福佬民謠	0
4	10	creator	11	songwriter	13	林福裕	0
5	10	description	15	lyric	16		0
6	10	description	15	notes	25	無伴奏「同聲三部合唱」。另有「混聲四部合唱」作品..	0
7	10	description	15	racerecord	24	新竹高中合唱團，何碧玲指揮，徐岱毓伴奏，蘇森墉...	0
8	10	description	15	structure	19		0
9	10	description	15	technique	18	蘇森墉將此曲分別編成「同聲三部合唱」與「混聲四...	0
10	10	extent	8	dimension	10	面	0
11	10	extent	8	quantity	9	4	0
12	10	format	7	extent	8		0
13	10	identifier	27	classificationNumber	28	S002005	0
14	10	identifier	27	uri	29	S002005_1.jpg	0
15	10	identifier	27	uri	30	S002005_2.jpg	0
16	10	identifier	27	uri	31	S002005_3.jpg	0
17	10	identifier	27	uri	32	S002005_4.jpg	0
18	10	lyric	16	contents	17	天烏烏要落雨，阿公仔拿鋤頭要掘芋。掘啦掘！掘啦掘..	0
19	10	record_list	0	creator	11		0
20	10	record_list	0	description	15		0
21	10	record_list	0	format	7		0
22	10	record_list	0	identifier	27		0
23	10	record_list	0	language	26	中文	0
24	10	record_list	0	title	5		0
25	10	record_list	0	type	1		0
26	10	structure	19	key	20	羽調式	0
27	10	structure	19	meter	21	2/4拍號	0
28	10	structure	19	musicalForm	23	三段式	0
29	10	structure	19	tempo	22	Moderato(中板)	0
30	10	title	5	mainTitle	6	天烏烏	0
31	10	type	1	classification	3	同聲三部合唱	0
32	10	type	1	localLevel	2	樂譜	0
33	10	type	1	style	4	合唱曲	0

圖 4- 2：單筆 Metadata 記錄格式內容範例

實作上，我們採用二層式資料結構(如圖 4- 2)，並賦予原文件中每一 XML 標記一特定 ID 值，當資料載入資料庫時做為唯一的識別鍵。使用者透過簡單的 SQL 指令，指定 XML 標記名稱與 Metadata 名稱，即可載出所需標記內的資料內容，而不需將原本 XML 內容資料載出再進行操作，增進系統效率及減少負擔。

又因 XML 文件本身需為 Well-Format 格式(亦即所有的標記不得交互存在)，為一巢狀結構。由於本系統採用 XML 文件作為載入內容及內容產出，故在載入 XML 文件之前，需先載入 DTD 或 Schema 作為格式規範，並具以產生欄位對應規則，以驗證資料內容格式是否統一。藉由 DTD 或 Schema 來驗證資料，可以統一文件格式並提高文件的可再利用性。

第二節主題架構

本研究的主題在於讓使用者依據本身的資訊需求，訂定個人化的知識組織架構來瀏覽所需的資訊，並將資料集(collection)自動置入於所自訂的知識分類架構中，並依關連性來建立相互間的鏈結架構。在雛型系統中我們提供使用者建置個人化知識鏈結的機制與知識組織相互分享的模組，讓使用者可依自己的需求來建立知識組織的架構。

系統初始階段藉由典藏資料中的 XML 之標記建立了分類組織架構。但不同的使用者對資料的組織瀏覽有不同需求，單一使用者對單一主題在不同情境下有的不同組織分類方式，而多個使用者對資料的觀點更有不同的看法，而且彼此的分享與引用他人已建立之知識分類架構，更可增加使用者個人瀏覽的需求。使用者分享其特定架構時，可設定分享的深度與廣度，在不同的情境下，提供不同的內容架構，讓使用者具有相當的彈性。

茲考慮相同的知識內容常因不同的人與觀點而有不同的需求或呈現方式，且人對於單一內容在不同情境下，亦可能存在多種的知識分類呈現方式，並且單一的知識內容常與不同類別的知識內容存在著關連性，所以，實作上我們讓使用者能操作知識間的關連性，可讓使用者將不同的類別連結起來，讓原有的階層式分層概念，擴展為網狀分類(networking)，完整的呈現使用者對知識組織鏈結的架構。而為達上述目的，在針對任何主題架構進行操作時，需以單一的節點為單位，從而考慮其上下位從屬關係、順序和連結屬性及使用者本身權限，則我們將節點物件之定義如圖 4- 3：

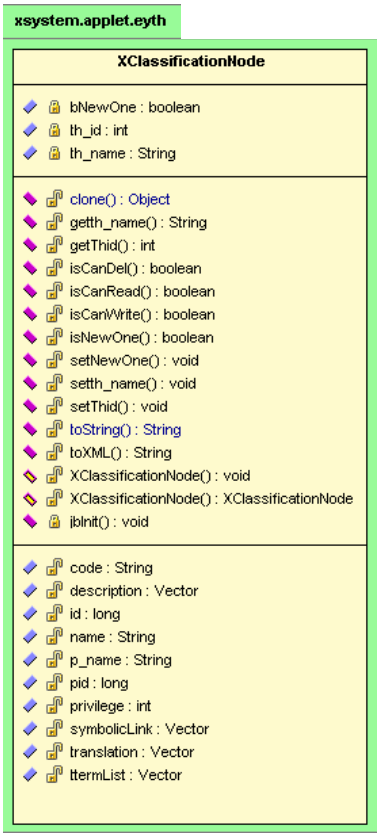

圖 4- 3：節點物件內容

茲說明該節點物件主要欄位功能定義與描述其記載內容；由於知識階層架構為一樹狀結構，故針對每一分類節點，皆需考量到其上下位關係與所屬層次之順序，故系統中用 pid 來記載其所屬父欄位編號，此欄位為一系統時間產出之長整數，作為此一主題架構中識別上層節點之資料。系統依據此一欄位內容，識別該節點所歸屬之上層節點，並由此可向上追溯，id 記載為本身之欄位編號，亦為一系統時間產出之長整數，作為此節點在此主題架構中識別鍵值使用。Description 記載欄位本身的說明，作為系統對外顯示供人閱讀之文字內容。另外，為達到最大的架構分享彈性，系統針對每一節點可賦予不同的操作權限，privilege 即記載權限內容，記載可予操作權限值。權限本身分為「讀取」、「寫入」和「刪除」，分由三個不同的數值所代表，如表 4- 1。彼此之間可以組合出不同的變化，供判斷使用，例如：權限值為 5 時，即表示擁有「讀取」與「刪除」之權限操作能力。屆時分享分類架構時，可依欄位值與使用者之權限比對，判斷其他使用者

對於此一節點可以進行那些操作內容。

表 4- 1：權限說明

數值	內容
1	讀取
2	寫入
4	刪除

為達到分類架構之間彼此之互相鏈結，採用 symbolicLink 來記載鏈結內容，記載與其他節點的鏈結狀況列舉項目，包含鏈結的屬性（參見或反見）。透過簡易查詢功能，可以幫助使用者搜尋相關節點，並決定是否與其建立關連。而實際鏈結的資料則是記載在 ttermList 之中，記載與 Metadata 資料實際鏈結資料內容。其中包含了關鍵字、欄位內容、Metadata 紀錄編號和所屬類別，如圖 4- 4：

	ONTO_ID	META	PARENT	TAG	ID	SEQ	TTERM
1	46	ntu_89_slik_dtd	slik	subject	1112926673437	0	文物
2	46	ntu_89_slik_dtd	slik	category	1112926710468	0	玄奘法師與大唐王朝(中國文物)
3	46	ntu_89_slik_dtd	slik	category	1112926757187	0	玄奘法師的中亞足跡(中亞文物)
4	46	ntu_89_slik_dtd	slik	category	1112926763093	0	玄奘法師朝聖之地印度(印度文物)
5	46	ntu_89_slik_dtd	slik	category	1112926764593	0	傳說中的玄奘法師及西遊記的世界
6	46	ntu_89_slik_dtd	slik	subject	1112926674515	0	石窟
7	46	ntu_89_slik_dtd	slik	category	1112926712531	0	甘肅天水，麥積山石窟
8	46	ntu_89_slik_dtd	slik	category	1112926713984	0	甘肅蘭州，炳靈寺石窟
9	46	ntu_89_slik_dtd	slik	category	1113442869750	0	甘肅張掖，馬蹄寺
10	46	ntu_89_slik_dtd	slik	category	1113442879968	0	甘肅酒泉，魏晉古墓群和畫像磚
11	46	ntu_89_slik_dtd	slik	category	1113442880921	0	甘肅安西，榆林窟
12	46	ntu_89_slik_dtd	slik	category	1113442881906	0	新疆吐魯番，吐峪溝石窟
13	46	ntu_89_slik_dtd	slik	category	1113442881906	1	新疆吐魯番 柏孜克里克石窟
14	46	ntu_89_slik_dtd	slik	category	1113442883062	0	新疆焉耆，明屋
15	46	ntu_89_slik_dtd	slik	category	1113442884296	0	新疆庫車，森木撒姆石窟
16	46	ntu_89_slik_dtd	slik	category	1113442884296	1	新疆庫車，庫木吐拉石窟
17	46	ntu_89_slik_dtd	slik	category	1113442884296	2	新疆庫車，克孜爾噶哈石窟
18	46	ntu_89_slik_dtd	slik	category	1113442884296	3	新疆庫車，克孜爾石窟
19	46	ntu_89_slik_dtd	slik	category	1113442885328	0	阿富汗 巴美揚 (梵衍那)石窟
20	46	ntu_89_slik_dtd	slik	subject	1112926676250	0	壁畫
21	46	ntu_89_slik_dtd	slik	category	1113442888468	0	印度 阿旃多（阿姜他）石窟
22	46	ntu_89_slik_dtd	slik	category	1113443096546	0	沙門守戒自殺緣品
23	46	ntu_89_slik_dtd	slik	category	1113443098218	0	九色鹿王本生
24	46	ntu_89_slik_dtd	slik	category	1113443099750	0	須摩提女緣品
25	46	ntu_89_slik_dtd	slik	category	1113443101484	0	須達拏太子本生
26	46	ntu_89_slik_dtd	slik	category	1113443102875	0	得眼林

圖 4- 4：ttermList 記載內容示例

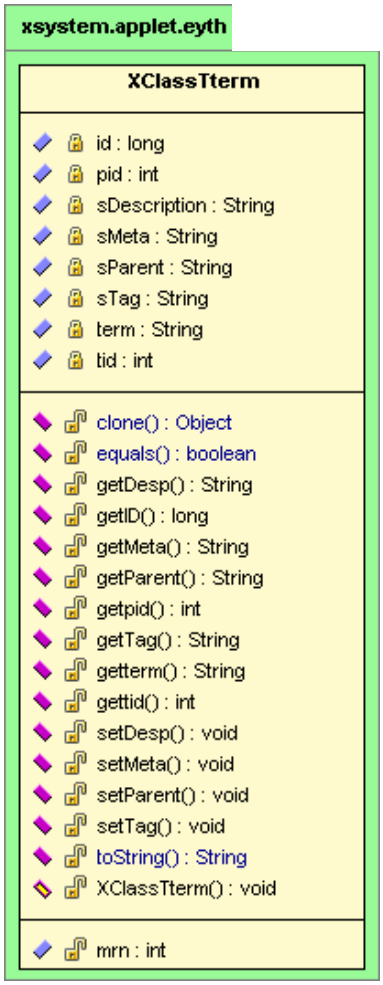

圖 4- 5：ttermList 物件

ttermList 物件(圖 4- 5)主要之記載內容有 id 用以識別其所屬之分類節點，sMeta 則是記載所屬 Metadata 代碼，用以區分不同的 Metadata，最後採用 term 記載鏈結的關鍵詞，記錄該筆 Metadata 記錄所採用之關鍵詞。系統將採用此欄位資料，對未分類資料內容，進行自動分類，以減少人工操作的時間，並增進效率，mrn 則是該筆記錄之系統識別號，作為 Metadata 記錄之間識別用。

經由物件的定義，詳細定義彼此之間操作方式(Method)與物件屬性，將複雜的過程隱藏於其內，簡化設計本身的流程與程式的開發。

第三節 文件自動分類

面對者每天不斷新產生的數位文件，要依賴人工來進行文件的整理分類，顯得困難且不可行，所以如何透過對現存已建置分類的資料內容(如國家型數位典藏計畫中已分類之 XML 資料)進行訓練分析，將統計分析的結果用於對新增或未分類的資料進行自動分類，來減少使用者人工分類的時間，增加系統擴充性。文中透過向量空間的文件特徵值運算，將已經由使用者分類的文件與未分類文件，藉由餘弦函數(cosine measure)計算其兩兩之間相似度平均值，再判斷未分類文件與已分類文件之間最小相似值作為比較，判斷未分類文件是否可歸於該類別。

系統由文件中關鍵詞彙所構成的特徵向量來計算文件相互間的相似度。假設向量 w_i 為文件 d_i 之內容，則文件 d_i 之內容以 $(w_{i1}, w_{i2}, w_{i3}, ..., w_{in})$ 表示，則具有 k 篇文件的分類項目 c 以下列符號表示：

$$c(d_1, d_2, d_3, ..., d_k)$$

其中若 d_j $(w_{j1}, w_{j2}, w_{j3}, ..., w_{jn})$ 為已知分類項目中之一篇文件，而 d_l $(w_{l1}, w_{l2}, w_{l3}, ..., w_{\ln})$ 為待分類之文件；而不同的關鍵字代表不同的維度，但每一關鍵字 w_{ik} 會因為出現在文章中位置和次數的不同而有不同的重要性，所以針對每一關鍵字系統會給予一個權重，如 d_1 $(w_1t_1, w_2t_2, w_3t_3, ..., w_nt_m)$。在本式中，$d_1$ 為已知分類項目中之單一文件，文字中之關鍵詞向量為 $(w_1, w_2, w_3, ..., w_n)$，而 t_1、t_2、t_3、...、t_n 代表這幾個關鍵詞 $(w_1, w_2, w_3, ..., w_n)$ 在 d_1 分類中，所代表之權重。

而權重的取得方式，則依關鍵詞在該篇文章的出現次數與出現在所有文章篇數倒數乘積 $(tf \times idf)$ 配合其所在的 XML 標籤加權而得到，亦即其權重越高，越能代表該篇文件。為此，本研究採用共同訊息向量法(mutual information vector)[2]，將 XML 文件中之關鍵詞與文件之關係，轉換為向量矩陣，再進行運算。其公式如下：

$$mi_{ef} = \log \frac{\frac{c_{ef}}{N}}{\frac{\sum_i c_{if}}{N} \times \frac{\sum_j c_{ej}}{N}}$$ ，其中 $N = \sum_i \sum_j c_{ij}$ 。[3] (1.1)

其中 N 為文件篇數，c_{ef} 為特定關鍵詞 f 在文件 e 中的出現次數，c_{if} 為文件 e 中的總詞數，c_{ej} 為該詞的總出現次數。而透過此方程式的轉換後，利用餘弦函數求取任意二向量 e_i 與 e_j 之夾角：

$$sim(e_i, e_j) = \frac{\sum_f mi_{e_i f} \times mi_{e_j f}}{\sqrt{\sum_f mi_{e_i f}^{\ 2} \times \sum_f mi_{e_j f}^{\ 2}}} \quad (1.2)$$

即可得二者之相似度。

系統中，針對其類別已分類文件集 C，由餘弦函數求取分類內容兩兩之間相似值，取其前 25%篇平均相似值 $\overline{m}$ ，再與欲分類之文件 d_j ，個別作餘弦函數運算，求取最小相似值 n，若 $n < \overline{m}$ ，則將文件視為該類別。

針對自動分類的結果，系統允許使用者對於文件分類的結果進行調整，對於不符條件文件重新歸類，降低語意上的誤差產生，透過使用者手動調整改善整體精確率。

第四節個人化知識展現架構

針對單一的知識主題項目，系統提供可針對多項分類項目進行鏈結，完整呈現其原有的多樣化內容，並針對多種主題，亦提供多個鏈結方式，以建立起各主題間的多樣化關連。對於鏈結的處理與呈現方式，有以下幾種方式；

一、單一 Metadata 多種呈現方式

使用者可以針對單一類型 Metadata 以多個主題面向來呈現。也就是說使用者可以從不同的觀點，對同一群 Metadata 進行不同分類主題階層架構的知識組織(如圖 4- 6 及圖 4- 7)。使用者在建構個人化的知識樹時，可以直接引用修改現存已建立的分類樹，而不需要全部重新建構。

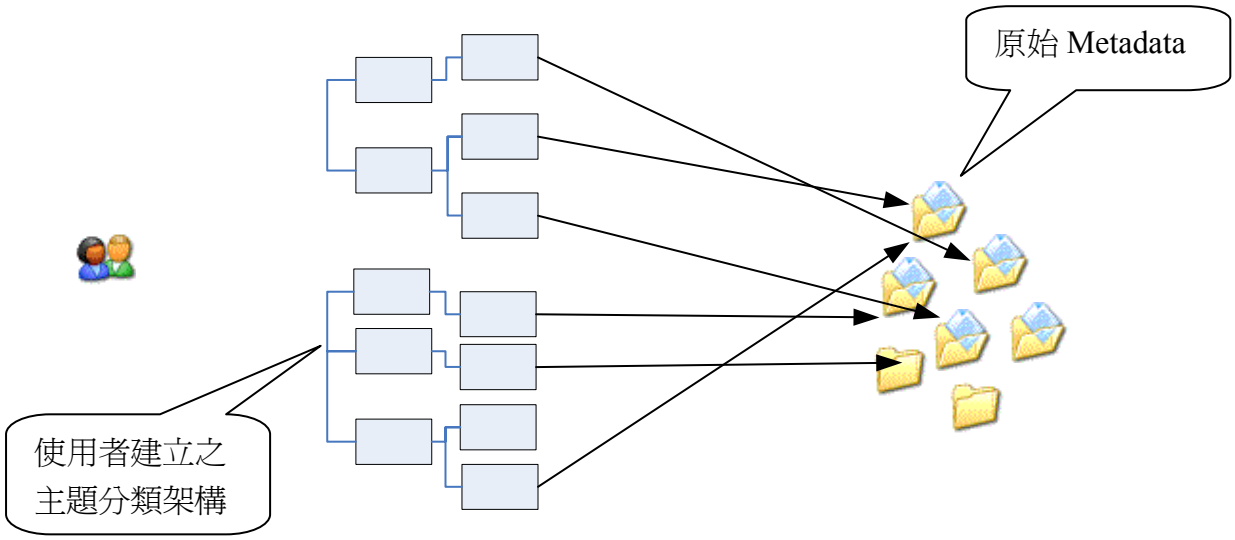

圖 4- 6：單一 Metadata 多種分類方式

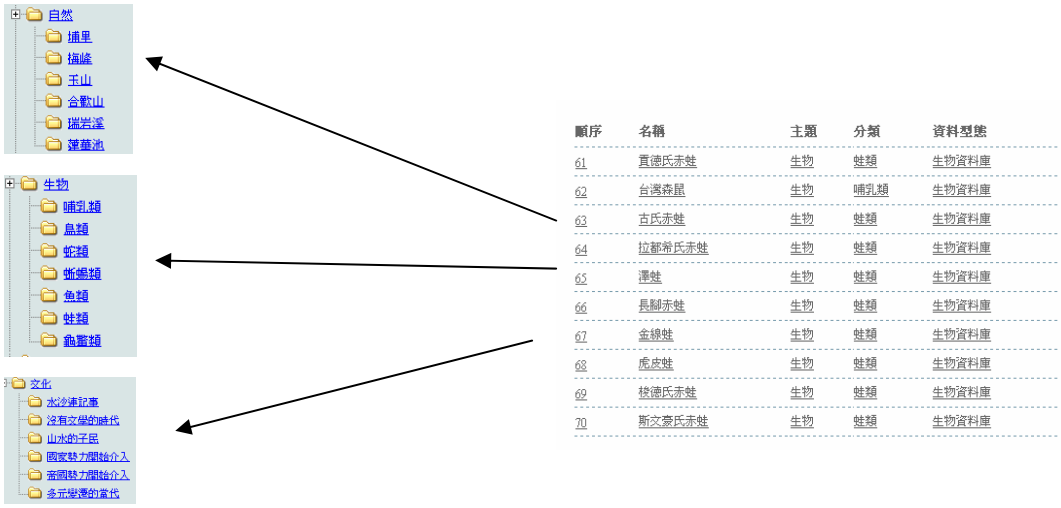

圖 4- 7：實際單一 Metadata 資料多種分類示例

二、多重 Metadata 多種呈現方式

系統中分類的資料並不侷限於單一類型的 Metadata，使用者亦可以針對多類型 Metadata，依使用者的知識組織架構來建立多種分類呈現方式，完整的展現彼此之間的從屬關係(如圖 4- 8)。

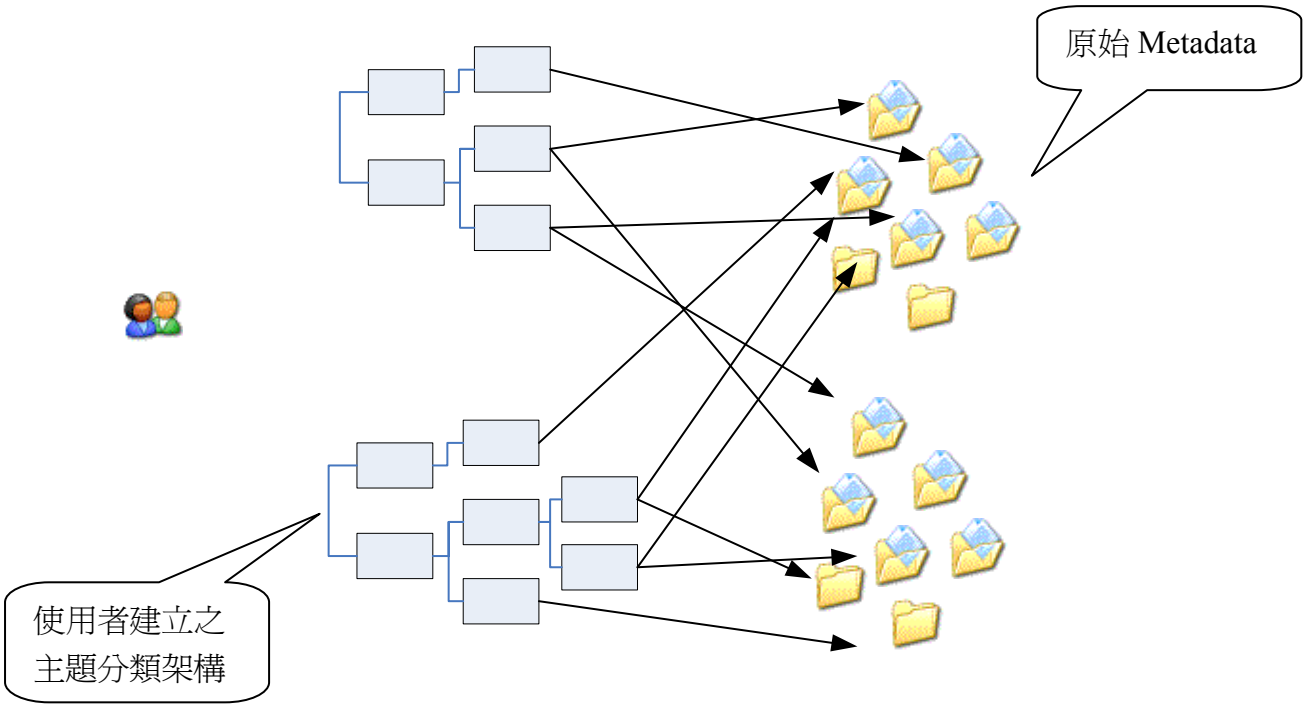

圖 4- 8：多種 Metadata 多種分類方式

三、主題架構之間的關連處理：

知識的分類並非為單一獨立的，其相互間鈈隱涵各式的相關性。系統提供鏈結的操作，透過簡單的查詢，讓使用者可以把相關的項目藉由鏈結節點，給予串連。除了提供本身分類本體內部的鏈結外，亦提供外部鏈結，鏈結外部分類主題架構(如圖 4- 9)。使用者亦可使用多個鏈結，完整的表示其相關資訊，擺除只能由階層來進行定義分類的觀點，由階層式的分類，擴展成為網狀的串連。(如圖 4- 10)

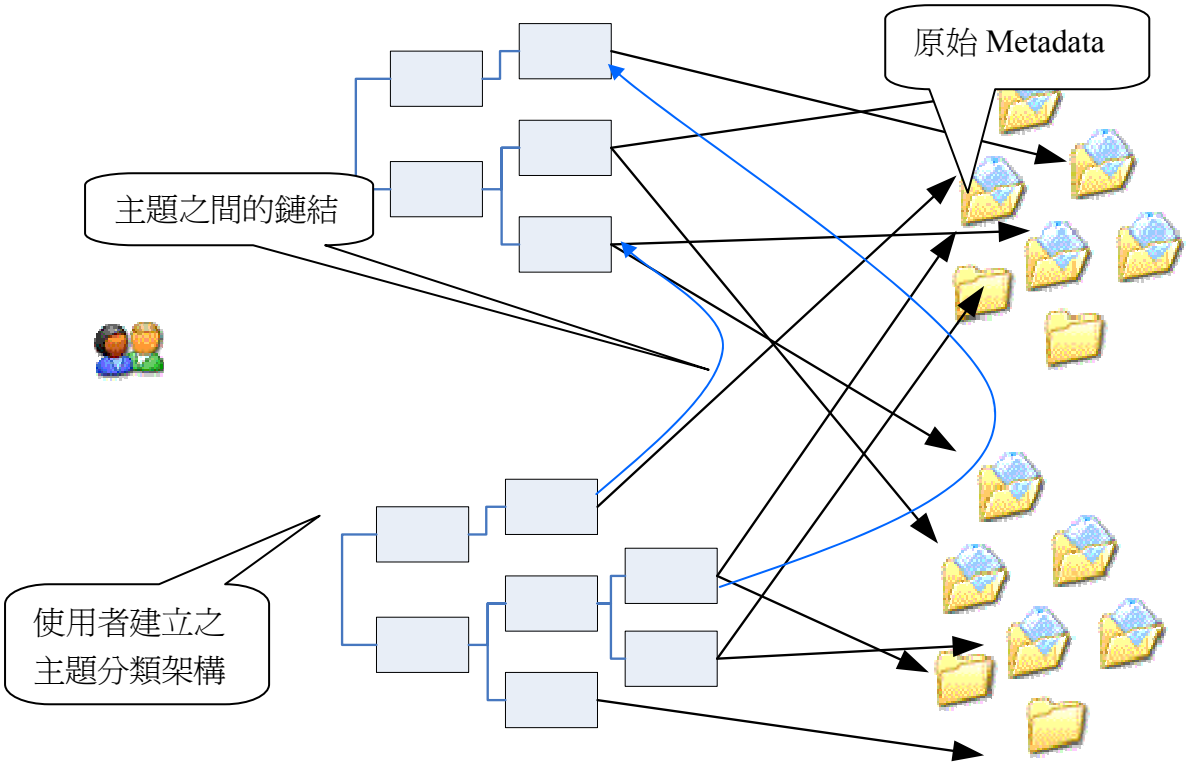

圖 4- 9：各主題樹之間的相關鏈結

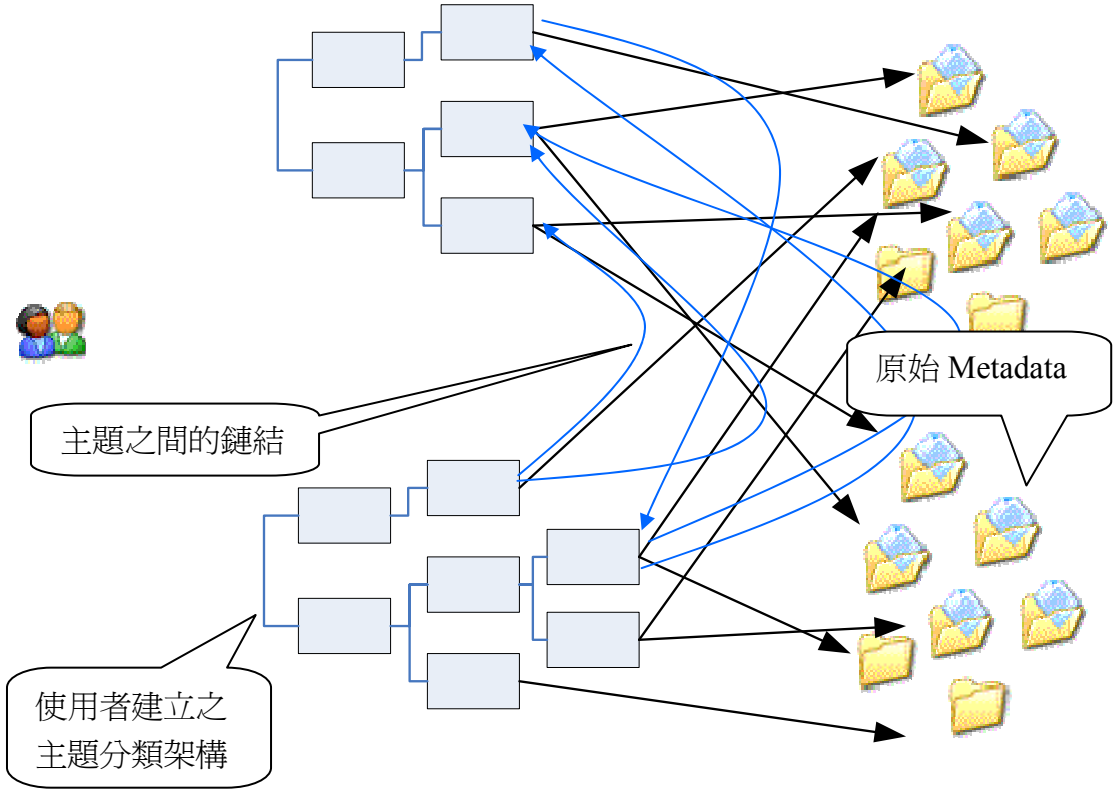

圖 4- 10：網狀的鏈結型態

第五節主題架構的引用

系統中允許使用者可以引用已經存在或別人所設定好的知識組織架構，或分享自己建置之知識組織架構，但依使用者個人之所屬的操作權限及欲引用之架構的權限設定內容，針對同一知識組織架構對不同的使用者而將有著不同的呈現結果。(圖 4- 11)

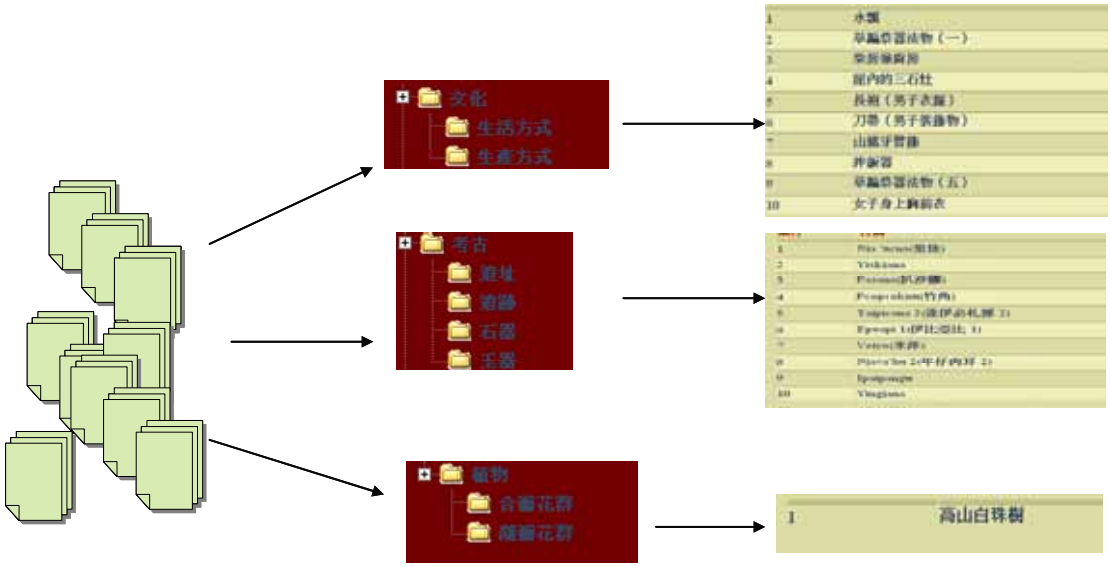

圖 4- 11：依據個別使用者不同的權限，相同的知識組織架構將呈現不同的內容

如第一節所述，使用者所分享的知識組織架構，針對其他使用者的呈現結果，是依使用者本身所具有的操作權限內容和主題架構本身所具有的權限來做判斷。

為了便於資料移轉及與外在系統的溝通、互動。本系統對外轉出與轉入資料格式皆為 Well-format 格式 XML 文件，並採用 UTF-8 編碼，以期能相容於各式字元，同時依據該主題架構的權限設定內容轉出內容。

[1] JDBC API Overview (http://java.sun.com/products/jdbc/overview.html)
[2] Pantel, P. and Lin, D. 2002. Document clustering with committees. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 199-206.
[3] 同註 2。

第五章系統測試

第一節測試資料輸入

本章中透過實際資料的測試，來檢驗系統是否有達成原先規劃的目的，讓使用者可以依個人化的需求來組織或架構所欲瀏覽的知識內容。

論文中我們採用國科會數位典藏計畫產出之 XML 文件做為測試資料(資料內容範例如圖 5- 1)，合計 9104 筆，其資料類別與資料數量分佈如表 5- 1。由於各類型資料所使用的標記定義與格式並不一致，無法依據其內容標記直接轉換成系統所需的欄位或抓取特定標記作為顯示區分，如圖 5- 2 所示，不同的 Metadata 其所採用的標記欄位不盡相同，針對 title 這一標記而言，可能在前一份文件被當作一主題欄位指出文件標題為「芟白荀」，但在另一份文件用於修飾主題內容，說明文件內容的描述為「蘭嶼昆蟲」、「蘭嶼蝴蝶典藏庫」、「小灰蝶科」等，所以需依據各不同主題類型的資料，截取其可代表內容的標記欄位，做為設定目錄及分類之用(表 5- 2)，並藉由涵蓋多個主題架構資料來驗證系統是否能針對不論單一主題的 Metadata 或涵蓋多主題資料，提供使用者依個人的需求來重新組織瀏覽所需的資訊。

```
<?xml version="1.0" encoding="UTF-8" ?>
- <historicalPhotos>
  <aggregationLevel>單件</aggregationLevel>
  <originalSurrogate>重製</originalSurrogate>
  <worksType>史料</worksType>
  <medium>數位檔案</medium>
  <quantity>1張</quantity>
  <dimension>原件為長12.5cmX寬8.9cm</dimension>
  <digitalCategory>影像(圖片)</digitalCategory>
  <fileName>cca100057-hm-hp_t_2003_01_0031-0001-u.tif</fileName>
  <fileName>cca100057-hm-hp_t_2003_01_0031-0001-i.jpg</fileName>
  <fileName>cca100057-hm-hp_t_2003_01_0031-0001-t.jpg</fileName>
  <color>彩色</color>
  <mainTitle>拉珏師當國禎</mainTitle>
  <subjectMatter>台灣陶瓷文化歷史照片</subjectMatter>
  <keywords>中華登龍陶瓷公司</keywords>
  <keywords>當國禎</keywords>
  <keywords>拉珏</keywords>
  <conditions>良好</conditions>
  <acquireMethod>任克重先生借入</acquireMethod>
  <creatorDynasty>民國</creatorDynasty>
  <contributor>任克重先生提供</contributor>
  <localFileName>cca100057-hm-hp_t_2003_01_0031-0001-u.tif</localFileName>
  <localFileName>cca100057-hm-hp_t_2003_01_0031-0001-i.jpg</localFileName>
  <localFileName>cca100057-hm-hp_t_2003_01_0031-0001-t.jpg</localFileName>
  <language>中文</language>
  <contentPlace>北投</contentPlace>
  <createdPlace>北投</createdPlace>
  <holderName>任克重</holderName>
  <rightsStatement>[20031101,]</rightsStatement>
  <accessRestrictions>本數位作品授權行政院文化建設委員會國家文化資料庫存檔及網際網路永久非商業性使用</accessRestrictions>
  <ownerCountry>中華民國</ownerCountry>
  <ownerName>本數位作品由台北縣立鶯歌陶瓷博物館及行政院文化建設委員會國家文化資料庫分別典藏,「原件由任克重先生典藏」</ownerName>
</historicalPhotos>
```

圖 5- 1：轉入系統之 XML 文件範例

```
<?xml version="1.0" encoding="big5" standalone="yes" ?>
- <pages>
  - <page>
      <pageid>2</pageid>
      <author />
      <title>埔里盆地兩棲類簡介</title>
      <content>　　埔里地區兩棲類十分豐富，蛙類主要以農田、沼澤、茭白筍田、空心菜
        區，蛙類常利用潮濕的樹林灌叢、溪流等自然環境以及人工之果園、稻田園為棲息地
        調查，結果共發現19種，約佔台灣所有蛙類物種之2/3，真可謂是蛙類的天堂。 <
        extension="jpg"/></content>
    - <metadata>
        <name>館別</name>
        <value>自然館</value>
      </metadata>
    - <childmedia>
        <childmid>292</childmid>
        <sequence>1</sequence>
        <title>茭白筍</title>
        <author>謝宗宇</author>
        <provider>謝宗宇</provider>
        <extension>jpg</extension>
        <description />
        <material>正片</material>
        <source>正片掃描</source>
        <time />
        <position>埔里</position>
      </childmedia>
    </page>
```

```
<?xml version="1.0" encoding="utf-16" ?>
- <exportContent version="1.0">
    <subject title="蘭嶼昆蟲" level="0" />
    <subject title="蘭嶼蝴蝶典藏庫" level="1" />
    <subject title="小灰蝶科" level="2" />
  - <item type="text">
      <![CDATA[　　角紋小灰蝶之分類地位為鱗翅目 (Lepidoptera) 小灰蝶科 (Ly
      動於寄主植物附近。本種亦廣泛分布於台灣全島平地到低山地區，海拔100-500 公尺
      ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2667_007.jpg ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2666_006.jpg ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2663_003.jpg ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2671_011.jpg ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2662_002.jpg ]]>
    </item>
  - <item type="image">
      <![CDATA[ images/image2670_010.jpg ]]>
    </item>
```

圖 5- 2：二相異 Metadata 格式示例

表 5- 1：各類型資料筆數

Metadata 類型	筆數
人體奧秘展覽館	237
大埔里人文與自然	531
玄奘西域行	278
宋詞三百首	310
阿里山山脈與鄒文化	523
唐詩三百首	321
荔鏡記	55
基因密碼展覽館	48
掌中布袋戲之人物篇	887
掌中布袋戲之樂器、兵器、道具篇	84
掌中布袋戲之頭戴、服飾篇	1930
楊英風之雕塑篇	160
楊英風之繪畫篇	956
蓬萊淨土遊	577
蝴蝶生態面面觀	354
蘇森墉音樂館	53
蘇軾詩詞	2854
蘭嶼生物與文化	976
合計	9104

表 5- 2：各類型 Metadata 之主題欄位

Metadata 類型	標記欄位
人體奧秘展覽館	Descriptions、subject、title、uri
大埔里人文與自然	Subject、location、familyname
玄奘西域行	Category、subject
宋詞三百首	Title
阿里山山脈與鄒文化	Subject、title3、type
唐詩三百首	Poem_Style
荔鏡記	LM_Chapter
基因密碼展覽館	Subject、title
掌中布袋戲之人物篇	mainSubject

掌中布袋戲之樂器、兵器、道具篇	localLevel、mainSubject
掌中布袋戲之頭戴、服飾篇	localLevel、mainSubject
楊英風之雕塑篇	primarySubject、styleAndMovement
楊英風之繪畫篇	primarySubject、styleAndMovement
蓬萊淨土遊	Category、subject
蝴蝶生態面面觀	Cfamily
蘇森墉音樂館	localLevel、style
蘇軾詩詞	SuShiPoem_VolumeName
蘭嶼生物與文化	Catalogy、subject、sub_subject

以楊英風之雕塑資料篇為例，使用者可借資料類別特性與 XML 文件的標記架構來建構個人化知識組織瀏覽模式，並將其內容資料組織與設定分派至自訂之知識類別。在圖 5- 3 中，區塊 A 為已存在之知識組織階層架構，區塊 B 為設定鏈結至各相關節點，作為各瀏覽階層之間互相參見鏈結之用，並可重複多個，區塊 C 為分至該類別之資料內容，區塊 D 為尚未分類之資料，並可透過「詞彙顯示」的功能設定更改成「顯示未分類詞彙」或「顯示全部」詞彙，允許針對同一詞彙進行多種分類。使用者可以將欲分至該類別之詞彙，選至該類別，並可建立相關鏈結，區塊 A 中已建立鏈結節點，將呈現為藍色，同時按下「+/-」可建立多個鏈結設定(圖 5- 4)，設定節點與節點之間的鏈結，充分還原知識之間的脈絡關係，讓原有單純的階層式的分類架構，擴展成為網絡狀。

透過上述相關主題項目鏈結的操作，系統會以樹狀階層架構 (如圖 5- 5)作為呈現個人化知識瀏覽介面。使用者點選所要瀏覽的類別，系統會將類別與其所有內容和相關的鏈結一併呈現，包含瀏覽記錄的資料及其所設定相關分類階層，並提供多樣化的呈現方式(簡略、詳細與圖形瀏覽顯示)。使用者可以在階層中進行關聯參照鏈結，瀏覽上下位節點和點擊單一筆記錄進行文件內容查看，亦可以藉由文件內容提供的相關鏈結「跳躍」至其他節點進行瀏覽。

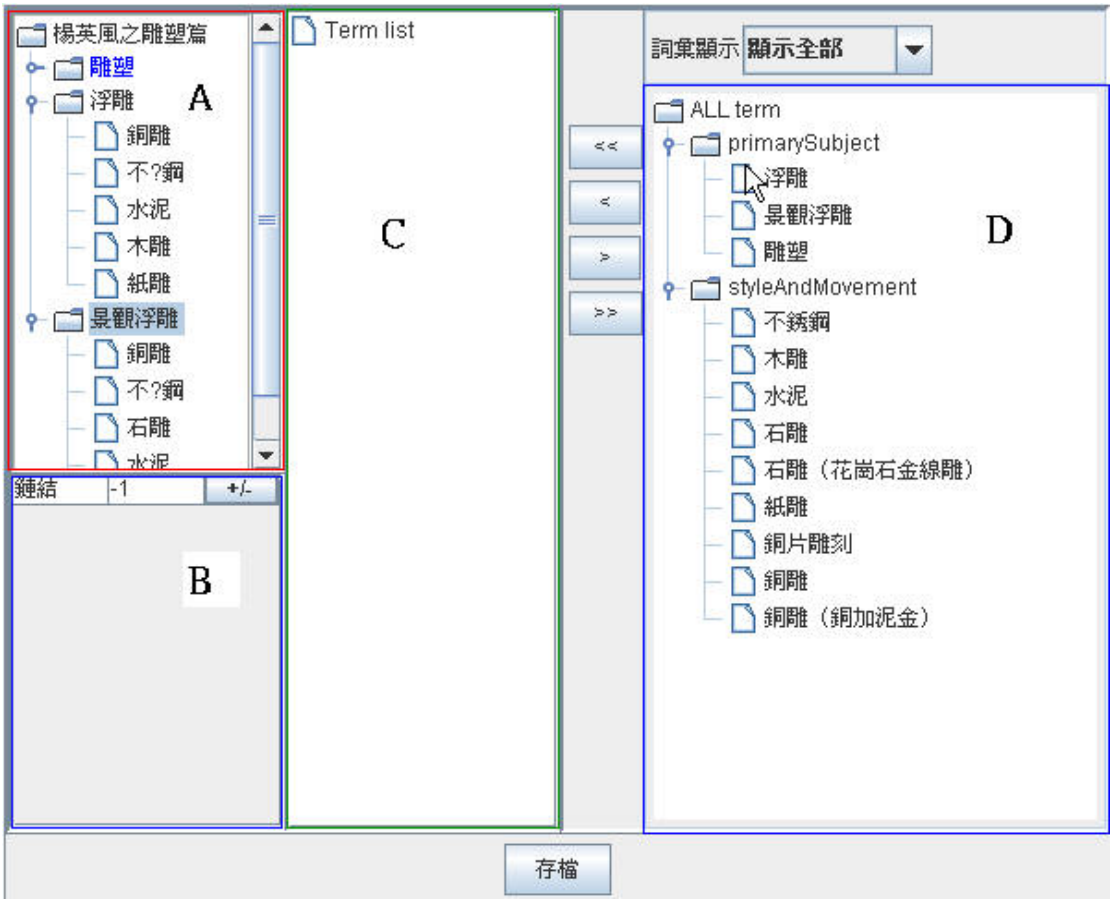

圖 5- 3：Metadata 資料分類設定畫面

圖 5- 4：相關鏈結設定畫面

圖 5- 5：結果呈現畫面

第二節 自動分類成效

由於數位典藏資料來源不斷的產生與更新，所以使用者針對現有項目分類完畢之後，一段時間後可能又有新的典藏資料產生要滙入系統，所以系統必須針對新增資料進行分類設定；或因資料內容龐大，全由人工進行分類設定，是一個相當耗時費力的工作，為此，系統引入自動分類的方式來降低人力的負擔。系統將已分類或部份採用人工分類的典藏資料作為訓練文件來建立自動分類統計門檻值，來對未分類文件或新進資料，進行相似度比對，把符合門檻值的項目置入該類。

使用者針對每一分類項目內容定義，可能存在相當大的差異性。故傳統針對單一分類給定單一固定特徵值（或稱門檻值）將變的不具有機動性與彈性。本研究採用單一分類項目中兩兩文件相似度的平均值，作為欲分類文件是否納入該分類之依據。如表 5- 3 為分類項目 A－不鏽網與分類項目 B－木雕部份相似值，但此二分類為各別獨立事件，亦即針對分類項目 A 或針對分類項目 B 所求取之特

徵並無法適用於另一分類項目。

表 5- 4：分類項目 A(不鏽綱)與分類項目 B(木雕)之相似值(節錄)

項次	A：不鏽綱	B：木雕
1	0.92590944445368	0.85386340964567
2	0.92025083154536	0.16052102851887
3	0.91478211304709	0.08078920524301
4	0.36533434041940	0.07941044718501
5	0.22655210607558	0.07187711246266
6	0.16621636118188	0.06093550849646
7	0.13501691067067	0.05990294147984
8	0.12857793657653	0.04505913543608
9	0.09255110561278	0.04157954904661
10	0.08642141568887	0.03808465775613
11	0.04978662246824	0.02298204869640
12	0.04329350380422	0.01881830877625
13	0.04329350380422	0.01842227120700
14	0.03975613543522	0.01380364851041
15	0.03293209292949	0.01294797275483
16	0.03182821782660	0.01167504963630
17	0.02548588012880	0.00294754220327
18	0.01573253163389	0.00106863559631
19	0.01375852155858	0.00096778099794
20	0.00431212420306	0.07187711246266
附註：依式 1.1.所求出之相似值內容，各取前 20 項目相似值。		

針對類別文件特徵值的計算，若採計所有分類中所有文件徵向量為平均相似度值做為該分類之門檻，則可能因取樣寬鬆而發生分類上的誤差，而導致自動分類成效過低，實驗中分別採計分類文件相似度平均值在前 25%、50%、75%與全部文件之平均相似值作為自動分類之依據，經初步實驗的結果，採用分類項目中之前 25%文件相似度平均值，做為該分類文件之特徵值門檻，此時文件的自動分類可以獲致較好的結果。但隨著類別文件的遞增，也許須要修正更嚴格的門檻值，對於分類的結果應該更加精準。

表 5- 5：單一分類項目中前 25%、50%、75%與全部文件之平均相似值

文件數量	平均相似值
25%	0.125337581124016
50%	0.032776177131946
75%	0.012680652318049
100%	0.003532560902032

系統中所採用 XML 文件資料格式標籤經常帶有豐富的語意訊息，和一般純文字文件有所區隔。實驗中我們引藉國家型數位典藏計畫所產生出的 XML 文件，做為實驗內容，其所採用的文件標籤，大多皆能描述其標籤內容。所以，除了採用文件本身之關鍵字、詞做為特徵值向量運算外，亦針對關鍵字詞在文件出現的位置之 XML 標記進行加權，再進行自動分類實驗。故原有之相似值公式除了原有之餘弦相似值外，另加入針對 XML 標記的權重，故針對分類向量向量e_i與e_j之夾角修正如下：

$$sim(e_i, e_j) = w_1 T + w_2 \cos(e_i, e_j)$$

其中 T 為 XML 標記相似值，w_1為針對 XML 標記所加的權重，而w_2為原有二向量餘弦函數所求之相似度。而 T 值的求法為一倒數之總和，若T_1為項目 A 中之標記，表示如下/a/b/c/d(即 b 為 a 之子欄位，而 c 為 b 之子欄位，以此類推)，T_2為項目 B 中之標記，表示如下/a/b/f/d，則可表示如下：

A：a/b/c/d

B：a/b/f/d

符合內容：1/1/0/0

則

$T = 1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3} + 0\times 2^{-4} = 0.75$

依據實驗所得之結果，發現以對 XML 標記增加權數 0.6 之分類效果最好，而加權之量越大，效果反而是大打折扣，而加權量越小效果只是不明顯，這可能是因為原始文件中 XML 所帶有的訊息已非常強烈，故太大的權重，反而會蓋過使用者原有分類集所帶有的訊息。

透過雛型系統的建置，驗證了論文中所提出架構模型的可行性，在此架構下，使用者自行定義組織與個人知識瀏覽架構模型的可行性，透過這樣一個系統，實現跨 Metadata 資料格式的操作及多樣化的瀏覽階層展現及分享。

第六章結論與建議

第一節結論

本論文提出一個人系統化知識組織與分享的方式，並經由第五章以資料實際測試，模擬使用者的操作狀況與資料儲存方式，透過雛型系統的建置與驗證，確認本研究所建立的架構是可行的。但本研究一開始即定位在以觀念取向的研究設計，因此在實際運作上，仍有相當多的細節需要加以規劃與考慮。例如：在瀏覽結構的呈現與執行速度的考量上，仍需調整至使用者能接受的時間範圍內，特別是針對大量資料所進行之壓力測試。又若欲建立自動化系統索引，以加速整體速度上，需調整各式索引結構與群集演算法，以達到更加精確與迅速之結果。

整體而言，本研究所獲得之成果如下：

一、 本研究架構可供各使用者自行定義個人之分類主題架構，並可分享自已定義的分類主題，供予其他使用者使用，而所產出之架構可以以 XML 格式載出或載入外部主題架構，作為資源分享與交換的目標，不會受到限制。

二、 分類主題可不限定於單一主題，更可針對單一主題建立多種分類主題，或者是多個主題擁有多個分類架構，系統皆可達成，並且提供簡易的操作方式，供使用者使用。為了充分表示各分類主題之間的關連，系統提供使用者可以建立多個鏈結至多項節點，以充分表示主題之間關連。

三、 針對分類結果的呈現，系統亦提供簡易的 HTML 頁面，配合

JavaScript 元件，做為資料的呈現與展示，並相容於各式瀏覽器。GUI 的呈現介面可提供使用者視覺化的瀏覽方式，並可快速的顯示其複雜的階層，作一全面性的瀏覽。

第二節未來研究方向

最後，在後續的研究方向則說明如下：

一、 針對使用者介面部份，本研究將設定部份之複雜邏輯交予 Java Applet 處理，並實際一 XML 訊息交換的方式，取代原有之 JDBC Connection，但此一方式在效率與 UTF-8 字元集上部份字元，如「锈」字會有呈現處理上的問題，建議後續的發展，需針對此一問題作較深入之討論。

二、 研究中利用現存已分類資料作為訓練資料，並用以針對未分類資料進行群集(Cluster)演算，但其複雜度過高，以致於系統整體效率不好，在此建議未來改善此一方式，或採用複雜度較低之演算法，以增進效率。

三、 配合 RDF 與 TOPIC MAP 的發展，未來轉出轉入之 XML 需參考其標準定義，唯仍需其間互轉之對照表，未來系統可以提供輔助建立對照表，將使用者的負荷減至最低；另外，在節點的鏈結部份，則需再加入屬性描述鏈結二端之節點關係，以相容於 TOPIC MAP 語法。