智慧型文件與智慧型系統整合之研究

林信成

副教授
淡江大學資訊與圖書館學系

摘要

本文首先從系統智慧化與文件智慧化兩個領域逐漸匯流的角度切入，探討電子文件的智慧化程度實是影響智慧型檢索系統性能的重要因素；接著，從XML發展趨勢觀之，我們認爲以XML爲核心的技術已經逐漸扮演了提升文件智能的重要角色；再者，藉由一系列的實作，我們以XML爲核心，建置了XML資料交換系統、XML新聞管理與出版系統、XML/CMARC編目系統和WAPOPAC行動公用目錄系統，分別驗證了XML在資料交換方面、在內文語意描述方面、在圖書館自動化的編目系統方面和行動資訊檢索方面，皆有著不可忽視的應用潛力。

關鍵詞：可擴展標示語言，智慧型出版，智慧型文件，智慧型系統，行動線上公用目錄

緒　論

網路普及一方面為人們帶來極大的便利，另一方面卻造成資訊的氾濫！如今，許多未經妥善組織整理的垃圾文件在網路上四處流竄，進而造成嚴重的資訊公害。因此，在資訊服務無所不在的今日，人們須要面對的一個嚴肅課題是：當大量文件數位化、網路化之後，使用者如何在浩瀚的文件庫中，找到所要的資訊？資訊基礎建設所標榜的終極目標，是要使任何人能在任何時間、任何地點，皆能透過網路獲得所需的任何資訊或服務，那麼，提供使用者一個有效的資訊檢索機制，便成為產出電子文件時所應考慮的重要課題。

為了提供使用者一個有效檢索電子文件的機制，近幾年逐漸形成了兩個蓬勃發展的研究領域：(一)智慧型系統之研究：此領域主要強調系統智慧化；(二)智慧型文件之研究：此領域主要強調文件智慧化。這兩者非但不是互不相容，而且還相輔相成，逐漸匯流(convergent)成為一個整合性的研究領域，其最終目的無非為了讓使用者能在浩瀚的電子空間順利且精確的查找到所需資料。

本文首先探討系統智慧化與文件智慧化之發展，然後提出一個重要觀點：認為電子文件的智慧化程度實乃影響智慧型檢索系統性能的重要因素之一，因此這兩個領域必將逐漸匯流成為一個整合性的研究領域；接著，我們從可擴展標示語言(eXtensible Markup Language，簡稱XML)(註1)發展趨勢觀之，可以很清楚的發現，以XML為核心的技術已經扮演了提升文件智能的重要角色；最後，本研究藉由一系列系統化的模擬與實驗，驗證了以XML為基礎的智慧型文件，不論在資料交換、電子出版、資訊檢索、資訊服務等方面，都極具應用潛力與價值。

二、系統智慧化與文件智慧化之匯流

長久以來，資料或文件在資訊檢索系統中都一直扮演著「被動者」的角色，必須等著「被」程式(Program)處理或檢索。因此，程式或演算法(Algorithm)是否夠聰明，便直接決定了檢索系統的智能高低。但是，即使像人工智慧這樣強調系統智能的研究領域，也不免碰上光靠提升系統智能也無法解決的問題。因此，在不斷提升系統智能之外，若能加強文件或資料本身的智能，絕對有助於提升檢索系統的整體性能。

㈠ 系統智慧化

系統智慧化之研究領域主要以資訊檢索(Information Retrieval，簡稱IR)技術為基礎，強調藉由研發更智慧型的資訊檢索系統，進而提升檢索性能，以滿足使用者的資訊需求。資訊檢索系統的首要工作之一是文件分析(document analysis)。其目的在於抽取出足以描述文件的特徵(feature)；而當使用者輸入查詢條件後，則進行查詢分析(query analysis)，並將此查詢映射(mapping)至文件空間(document space)中，然後計算「查詢」與「文件」之間的「相似度」(similarity)而完成比對(matching)工作，最後得到檢索結果。在文件分析過程中所抽取出的文件特徵，是否具有足夠的代表性而能充分描述整份文件，對於整個檢索系統的效能有決定性的影響。一般常被用來評估檢索系統性能的指標有回現率(recall rate)和精確率(precision rate)(註2)，此兩者經常是無法兼得的。有些檢索系統回現率有餘而精確率不足，一個查詢動輒找到成千上百篇的文章，其中卻只有少數滿足使用者真正的資訊需求，反而給使用者帶來資訊過載的困擾；反之，若僅顧及精確率則又往往犧牲回現率。

資訊檢索技術歷經數十年的發展，累積了不少經驗與成果，如自動索引(indexing)技術(註3)、自動文件分類(automatic document classification)技術

註1 XML乃由全球資訊網協會(World Wide Web Consortium，簡稱W3C)所提出，在1998年2月10日成為建議規格(Recommendation)，詳細內容可參見<http：//www.w3c.org>。

註2 Robert R. Korfhage, *Information Storage and Retrieval* (New York：Wiley Computer Publishing, 1997), pp. 196-199.

註3 Gerard Salton, "A Comparison between Manual and Automatic Indexing Methods," *American Documentation*, 20：1 (1969)：61-71.

(註4)、全文檢索(full-text retrieval)技術(註5)、相關回饋(relevance feedback)(註6)技術、自然語言處理(natural language processing)技術及跨語資訊檢索(cross-language information retrieval)技術(註7)等；隨著資訊量的累積，資訊檢索的難度日益增加，近年來由於電腦科技的發展及各領域研究人員的投入，資訊檢索技術漸漸朝向智慧型系統方向發展，如資訊過濾(information filtering)、資訊擷取(information extraction)、資料挖掘(data mining)、智慧型代理人(intelligent agent)等；而資訊內容的多樣性、多媒體化則為現今電子文件的重要特性，無論文字、聲音、影像、圖片、視訊、動畫，都可能出現在數位化文件中。此一趨勢則直接刺激了多媒體資訊檢索(Multimedia Information Retrieval)技術(註8)的進展。如此一來，資訊檢索遂成了一個多元化的研究領域，Michael Lesk將資訊檢索技術的歷程，從1945年起以每十年為一個年代劃分，每個年代都有重要的突破與進展，是瞭解近代資訊檢索技術發展的重要文獻之一(註9)。

(二) 文件智慧化

文件智慧化之研究領域以智慧型文件(Intelligent Document)為核心，所謂智慧型文件意指在原始資料當中加註了額外的語意式描述資料者。此領域的研究強調知識組織(Knowledge Organization，簡稱KO)，著重Metadata(註10)的著錄，藉由對原始資料進行詮釋、標示等加值處理，而使文件本身更具語意層次的自我描述性，進而能更精確的檢索出所需資料。

一般而言，電子文件依其結構性可概分為三大類(註11)：

1. 完全結構化文件(Fully Structured Document)：結構性完整的文件稱為完全結構化文件。例如，在關連式資料庫(Relational Database)的關連表(Relation Table)中所儲存的每一筆記錄(record)都有明確的欄位(field)定義，每個欄位的資料型態(data type)和大小也都清清楚楚，這便是完全結構化文件；又如利用MARC所著錄的書目資料、用Dublin Core所描述的網路資源等，也都是屬於此類。

註4 K. A. Hamil & A. Zamora, "The Use of Titles for Automatic Document Classification," *Journal of American Society for Information Science,* 43：2 (1992)：130-148.

註5 Gerard Salton, *An Introduction to Modern Information Retrieval* (New York：McGraw-Hill, 1983).

註6 Gerard Salton & C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science,* 41(1990)：288-297.

註7 陳光華，〈超越資訊檢索的語言藩籬〉，大學圖書館，2：1(民國87年1月)：頁87-99。

註8 曾元顯，〈多媒體資訊檢索技術之探討〉，21世紀資訊科學與技術的展望國際學術研討會(民國85年9月。)

註9 Michael Lesk, "The Seven Ages of Information Retrieval," available at <http://www.lesk.com/mlesk/ages/ages.html> (20 Feb, 2003).

註10 Metadata是「用來描述資料的資料」(Data describes other data)或「關於資料的資料」(Data about data)，其譯名有「元資料」、「描述資料」、「詮釋資料」...等，並不統一，因此本文直接採用原文而不使用譯文。目前已發展成熟或正發展的metadata格式眾多，適用於不同領域及用途，詳細內容可參考：陳雪華，〈網路資源與Metadata之發展〉，圖書館學刊，12期(民國86年12月)：頁19-37。

註11 林信成，〈XML相關技術與下一代Web出版趨勢之研究〉，教育資料與圖書館學，37：2 (民國88年12月)：頁184-210，亦可得自〈http://mail.tku.edu.tw/sclin/research/pub/XMLWeb.htm〉(民國92年2月20日)。

2. 完全非結構化文件(Fully Unstructured Document)：毫無組織的資料或毫無結構可循的文件，稱為完全非結構化文件。例如，缺乏段落結構的全文資料、遙測資料、監控數據、聲音、影像資料 ..... 等，這一類型的文件在儲存時較為簡單，然而由於欠缺文件結構資訊，使得檢索技術相對困難。

3. 部分結構化文件(Partial Structured Document)：介於以上兩個極端之間的稱為部分結構化文件。例如，章節段落清晰的全文資料，在此類文件中的題名、作者、摘要、章節段落等都是屬於結構化資料，而文件中的本文則是屬於非結構化資料。

實際上，文件中的結構化資訊，經常是資訊檢索系統進行特徵抽取時非常重要的依據。在進行文件分析或檢索時，通常可以藉由文件中結構化資訊的輔助，簡化分析過程或提高檢索性能。舉例而言，當對文件進行自動關鍵字擷取(key-word extraction)時，題名中的詞彙就比一般本文中的詞彙來得重要許多。因此，加強文件的結構性，增加描述性的Metadata，對於簡化文件分析過程，提昇檢索精確率有極大的幫助。以此觀之，在發展智慧型資訊檢索技術之外，加強Metadata的著錄，提升文件的語意層次，使文件因具備自我描述性(Self-Description)而更智慧化，實為提昇檢索精確率的有效方案。Metadata是個極為普遍的概念，在我們的日常生活中，四處可見Metadata的蹤影，例如：{CPU型號， 記憶體大小，硬碟機容量.....}，可用來描述個人電腦的規格；而 {書名, 作者, 出版社..... } 可用來描述出版品資料。為了讓Metadata發揮更大的功效，於是人們開始制訂各種Metadata標準以供遵循，圖書館長期以來所沿用的機讀編目格式MARC，就是用來描述書目資料的Metadata標準。在網路盛行之後，為了因應既多且雜的電子文件，讓使用者都能盡快而且精確的找到所需資料，陸續被制訂出的Metadata標準也就愈來愈多。Metadata可以在不同的系統平台中實作，而Web又是目前最重要的電子出版平台，因此，如何在Web平台上實現Metadata，以提升Web文件的智慧性，進而使得智慧型檢索系統能更精確的查找到相關資料，便成了另一個重要的研究議題。

總之，智慧型系統與智慧型文件的相關技術與研究方法若能有效整合，必能提供使用者更便利的資訊服務。

三、以XML爲核心之智慧型Web出版

以現今的Web而言，HTML仍是發行電子文件的標準規格。然而，HTML著重於版面編排與外觀格式，對於文件結構的規範及內容語意的描述則乏善可陳；再者，因其不具備可擴展性，所以使用HTML著錄Metadata的成效不彰。XML的誕生正好提供了一個可行的解決方案，為Metadata的實作提供了一個基礎平台。不過，XML並不是被發展出來取代HTML的，而是用以彌補其不足之處。XML

自1998年日2月10正式標準(註12)發佈至今，歷經五年多的發展(註13)，已經成為一個陣容龐大的技術家族：DTD和XML Schema(註14)用以定義文件的結構；CSS(註15)和XSL/XSLT(註16)分別作為呈現文件版面和轉換文件格式之用；DOM(註17)是剖析文件時的標準物件模型；RDF(註18)則作為Metadata之整合框架；Nmaespace(註19)是一致性名稱識別機制；以及各種衍生的應用語言如MathML(註20)、SVG(註21)、PNG(註22)、SMIL(註23) ... 等，再加上由各個不同組織基於XML所發展出適用於各行各業的應用語言將近千種(註24)，真可謂族繁不及備載。至於近期XML的研究則逐漸朝向智慧型的Web語意網(Semantic Web)(註25)和Web知識體(Web Ontology)(註26)的方向發展，使得Web系統與文件皆愈來愈智慧化。

依據XML的特性，可歸納出以XML技術為核心的智慧型電子文件將具備如下之特色：

* 電子文件具備自我描述性
* 電子文件更能有效整合
* 電子文件更具結構性
* 電子文件具備內容和外觀分離原則
* 標注語言具備多樣性及可擴展性

綜上所述，XML不但能有效解決目前網路上電子文件的亂象，更有助於開創智慧型電子出版的新契機。

註12 W3C, "Extensible Markup Language(XML)," available at<http://www.w3.org/XML/>(20 Feb. 2003).
註13 W3C, "Happy Fifth Birthday to XML-10 February 2003," available at<http://www.w3.org/>(20 Feb. 2003).
註14 W3C, "XML Schema," available at<http://www.w3.org/XML/Schema>(20 Feb. 2003).
註15 W3C, "Cascading Style Sheets," available at<http://www.w3.org/Style/CSS/>(20 Feb. 2003).
註16 W3C, "The Extensible Stylesheet Language(XSL)," available at<http://www.w3.org/Style/XSL/> (20 Feb. 2003).
註17 W3C, "Document Object Model(DOM)," available at<http://www.w3.org/DOM/> (20 Feb. 2003).
註18 W3C, "Resource Description Framework(RDF)," available at<http://www.w3.org/RDF/> (20 Feb. 2003).
註19 W3C, "Namespaces in XML," available at<http://www.w3.org/TR/REC-xml-names/> (20 Feb. 2003).
註20 W3C, "W3C Math Home," available at<http://www.w3.org/Math/> (20 Feb. 2003).
註21 W3C, "Scalable Vector Graphics (SVG)," available at<http://www.w3.org/Graphics/SVG/> (20 Feb. 2003).
註22 W3C, "PNG (Portable Network Graphics)," available at<http://www.w3.org/Graphics/PNG/> (20 Feb. 2003).
註23 W3C, "Synchronized Multimedia," available at<http://www.w3.org/AudioVideo/> (20 Feb. 2003).
註24 XML.ORG, "Applying XML and Web Services Standards in Industry," available at<http://www.xml.org/> (20 Feb. 2003).
註25 W3C, "Semantic Web," available at<http://www.w3.org/2001/sw/>(20 Feb. 2003).
註26 W3C, "Web-Ontology (WebOnt) Working Group," <http://www.w3.org/2001/sw/WebOnt/>(20 Feb. 2003).

四、系統實作

為了印證上述的論點，本研究透過系統實作方式，完成如下若干個實驗系統：

(一)以XML為基礎之資料交換系統

(二)以XML為基礎之新聞管理與出版系統

(三)以XML與CMARC簡篇為基礎之編目系統

(四)圖書館行動線上公用目錄系統WAPOPAC

茲將以上四個實驗系統劃分為四個研究單元來加以說明。

(一)以XML爲基礎之資料交換系統(註27)

資料交換系統是由若干個分散於網路上的不同系統所組成，彼此透過相同的協定進行資料傳遞與共享。本研究所建立的交換系統架構如圖1所示，圖中以「系統A」及「系統B」代表兩個分散於網路上的系統，彼此透過XML進行資料的傳遞與交換。在本系統中我們完成了七大功能模組，其主要功能說明如下：

1. 資料檢索模組(Data Searching Module)：本模組為一聯合檢索機制，檢索結果將透過轉換模組及傳送模組以XML格式傳遞。

2. XML轉換模組(XML Transforming Module)：本模組負責將檢索出的資料依雙方共通之DTD規範轉換為有效的(Valid)XML格式。

3. 資料傳送模組(Data Transferring Module)：本模組負責將轉換模組所轉換後之XML資料傳送至對方之XML剖析模組，以做進一步之資料檢驗。

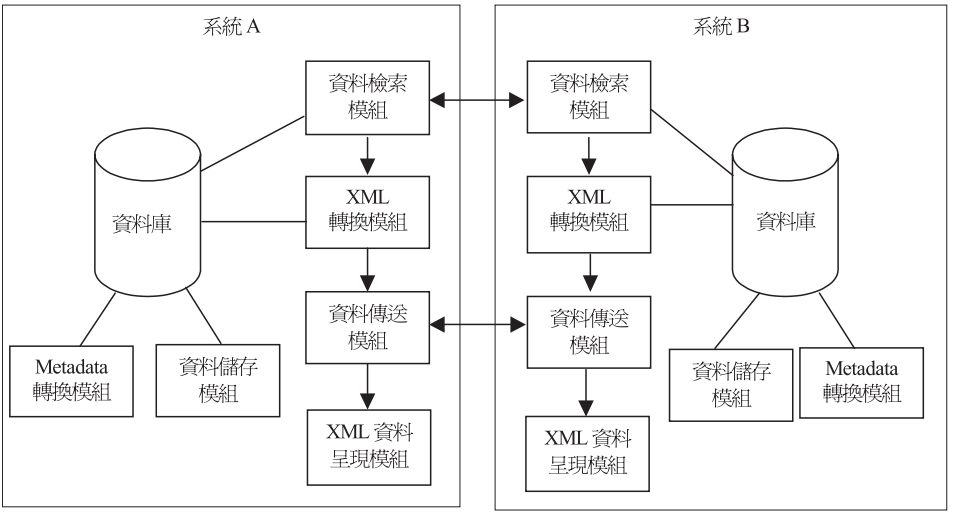

圖1 系統架構圖

4. XML剖析模組(XML Parsing Module)：本模組負責檢驗所交換的XML資料是否具有效性(Validation)，若此文件為有效的XML文件，則將其交由XML資

註27 林信成、陳勇任，〈基於XML之網際網路資料交換雛形系統設計〉，教育資料與圖書館學，39：2，(民國90年12月)：頁145-160。

料呈現模組輸出至使用者介面。

5. XML資料呈現模組(XML Data Presentation Module)：本模組首先透過DSO(註28)以及DOM(註29)來解讀XML文件，再依據資料所需之功能及超連結，加以包裝、排版，最後將資料內容呈現給使用者。

6. 資料儲存模組(Data Storage Module)：本模組能將解讀後的XML資料轉存回系統內之資料庫。

7. Metadata轉換模組(Metadata Transforming Module)：本模組內建Metadata對照表，可彈性的針對內部的Metadata做外部的Metadata轉換，並以XML格式輸出，提供了跨平台的資料交換的機制。

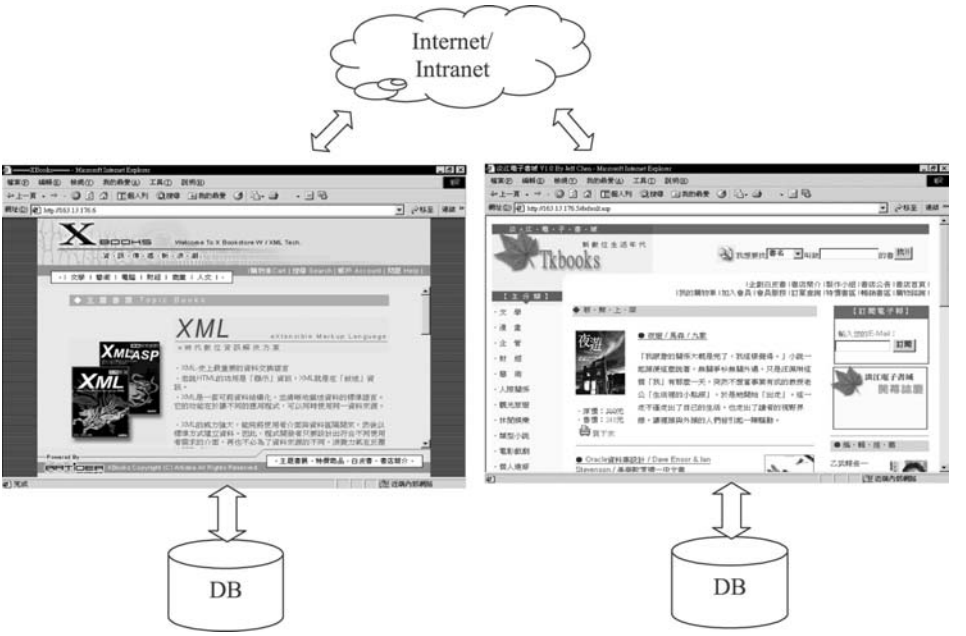

圖2 XML-Based資料交換系統

我們依據上述架構在Web上建置兩家虛擬書店，分別為代表系統A之「X電子書城」(XBooks)(註30)及代表系統B之「淡江電子書城」(TkBooks)(註31)，並在兩端分別植入前述七大模組作為資料交換及聯合檢索之用，如圖2所示。實驗結果各模組運作良好與前述分析相符。本系統針對XML資料的處理以及資料的交換做深入的探討，以發揮XML在資料處理以及資料交換的優勢。隨著各模組的成

註28 DSO(Data Source Object，資料來源物件)可將XML文件視為一份文件資料庫進行資料存取的動作。

註29 DOM(Document Object Module，文件物件模型)為W3C所制定的介面標準，定義了文件的邏輯結構和存取、處理、操縱文件的方法。DOM是一套普遍適用於HTML、XML等文件的應用程式介面(Application Programming Interface，API)。詳細內容可參考W3C網站<http://www.w3c.org/ DOM>。

註30 X電子書城(XBooks)之URL為<http://163.13.176.6>。

註31 淡江電子書城(TkBooks)之URL為<http://163.13.176.5>。

功開發，我們順利地建構了一個網際網路上的資料交換實驗系統。

(二) 以XML爲基礎之新聞管理與出版系統(註32)

接著為了驗證經由XML著錄Metadata之電子文件，能有效提升檢索系統之精確度，我們以電子新聞的管理與出版為例，藉由自訂的Metadata格式，以XML語法進行實驗性新聞資料庫之全文標誌，並自行設計一套新聞管理與出版系統，實際在Web環境中整合XML技術，測試加註了這些描述性資料之後的檢索結果，探討與印證XML在電子新聞管理與出版方面的優勢。此系統之特色為各個管理與出版模組皆以XML為基礎，系統內之所有資料亦採用XML格式，相較於傳統的資料處理模式來說，有著更彈性與更易加值處理的特點！再者，藉由XML優越的結構化與自我描述性，使得電子文件的「智慧化」程度得以提升，進而增進資訊檢索之精確度。而在新聞資料庫的Metadata方面，在參考國內外的新聞Metadata，包括政大謝瀛春教授發表的有關科學新聞的內容標誌(註33)，及國際上的兩大主流NITF(News Industry Text Format)(註34)與NewsML(註35)之後，由於考慮到本系統之實際需求乃在於提升內文檢索的精確度，因此我們自訂了一個適用於本系統的簡易版新聞Metadata DTD。其中，每則新聞除了包含諸如新聞編號、索引、日期、標題、作者..... 等基本資料，另將新聞內容分以人、事、時、地、物等加以標誌，以便進行內文語意搜尋之用，最後，尚含有一個排版用的標籤作為不同樣式之套版。

本系統之介面共分為前端使用者介面與後端管理者介面，若依功能劃分則可區分為四大功能模組，其主要功能說明如下：

1. 資料出版模組(Data Publishing Module)：本模組負責將資料內容呈現給使用者，透過DSO 以及DOM 來解讀XML文件，並結合該資料所需之相關功能及超連結，加以包裝、排版，只要是符合系統之DTD規範的XML文件，皆可透過此模組呈現內容。

2. 資料檢索模組(Data Searching Module)：本模組供使用者檢索所需新聞資料之用。一般檢索模組僅提供欄位的檢索，並未提供針對全文中某些特定對象的檢索，如人名、地名等，本模組除提供新聞類別、關鍵字詞檢索功能之外，藉由XML將文件全文標示，可針對新聞內容的人、事、時、地、物加以檢索，提高檢索結果的精確率。

3. 資料編輯模組(Data Editing Module)：本模組提供管理者編輯新聞文件內容。對於XML並不了解或不熟悉者，皆可透過此模組，輕易的將所需之新聞內容

註32 林信成、陳勇任、楊翔淳，〈基於XML之新聞管理與出版系統設計〉，2002出版與圖書館研討會，台北淡水，民國91年4月26日，頁14-29。

註33 謝瀛春、黃學碩、維習安、雷約翰、謝清俊，〈新聞內容的標誌-XML之應用〉，海峽兩岸資料庫/數據庫與資訊/信息服務交流與合作論文集(民國90年1月)，頁205-212。

註34 IPTC, "News Industry Text Format," available at<http://www.nitf.org/> (20 Feb. 2003).

註35 XMLNews.org, "XML and News Industry," available at<http://www.xmlnews.org/> (20 Feb. 2003).

編輯成符合系統之DTD規範的XML文件，並針對新聞內容給予人、事、時、地、物不同的標記，提供檢索模組使用，另外經由XML文件內容與呈現資料分離的特點，同一份文件可選擇不同樣式做為出版的選擇。

4. 資料管理模組(Data Management Module)：本模組提供管理者異動／修改資料。透過網路可遠端開啟管理模組，對所需的新聞資料做新增、修改與刪除等動作，另外，對於異動過的資料，藉由資料管理模組即可做查詢的動作，檢視其XML內容是否可正確呈現，無需回到一般使用者介面。

圖3 系統檢索結果

圖3為本系統之實驗結果，我們以「大學」作為檢索詞，在圖的左上方乃是選擇以不限標誌的方式進行內文檢索之結果，共找到八篇文章；對於同樣的檢索詞，如果將檢索條件限制於「地」，表示使用者想要查詢的只是與大學有關的地方、地名或地點，而非所有與「大學」概念相關的文章，結果如圖的右上方所示，共找到七篇文章；再者，若檢索詞仍為「大學」，但將檢索條件限制於「事」，表示使用者想要查詢的是有關大學的事件而非地點，則如圖的右下方所示，更精確的找到兩篇含有大學相關事件的文章。

本研究單元透過新聞管理與出版系統的實作，將各資料模組以XML為基礎，系統內所有資料亦採用XML格式，以XML將新聞作結構化的處理，並自訂Metadata來描述其內容，搭配資料檢索模組，可確實針對新聞內容作精確檢索，其精確度優於傳統的全文檢索結果。此外，由於XML資料與樣式分離的特性，使得新聞呈現的樣式非常有彈性，可因不同的使用者需求做更改，而不用更動原始的新聞資料內容。本系統之理念可應用在許多方面，以網路新聞為例，每家媒體有著不同的新聞格式與其排版方式，一但需要發佈在同一網站之上，必定另外制定合作的標準，才能解決彼此間不相容的問題；若運用本系統所提之原則，只需

遵守DTD的規範，同一份新聞內容，就可依照不同的需求，迅速方便的在網站上出版，而無需另外做內容上的更改。

㈢ 以XML與CMARC簡篇爲基礎之編目與檢索系統（註36）

我們接著以《中國機讀編目格式(CMARC)簡篇》第4版為標準，提出一份可行之DTD作為書目資料庫綱要，並透過系統實作方式以XML為基礎，設計一個編目模組，其目的在探討XML應用於圖書館自動化系統之潛力；此外，藉由系統實作，亦展現XML在結構化資料組織方面之優越性能。

CMARC長久以來一直是圖書館自動化系統所採用的編目標準之一，但由於國內自動化系統並不一致，所以往往在進行書目資料交換時會產生許多轉換上的問題。CMARC的漸趨複雜，也使得資料的建檔格外變得有選擇性，書目資料在線上的展現模式，不管以簡略(brief)或以簡短款目(short entry catalog)的方式，若以使用者的使用效度來考量，編目員建檔並不需要建立多完整或多完美的書目記錄才是唯一的編目目標，事實上讀者只需要一些有效的檢索要項，每筆資料也只要一些具可獲性的檢索點，這就足以幫助讀者找到資料並發揮線上目錄的功能（註37）。因此簡略編目對中小型及地方圖書館來說，已足以滿足讀者的需求，並且使得中小型自動化系統更容易開發，節省圖書館經費。

CMARC可視為Metadata的一種，它是描述書目資料的資料，而XML是實現Metadata的語法之一，要使Metadata具備互通性，可以在不同系統之間交換，則包裝Metadata的語法是非常重要的部份。猶如ISO 2709在不同的圖書館自動化系統之間穿梭自如，HTML是使得WWW文獻能在不同系統間交換的主要功臣，而SGML、XML是電子圖書館/博物館/檔案館系統用來標示其Metadata及全文資料的標準語法(註38)。本研究採用的《中國機讀編目格式》為第4版(以下簡稱*CMARC4*)，DTD的內容是以*CMARC4*簡篇(註39)為標準，再參考國家圖書館所制定之「NBINet核心書目記錄必備項目說明」中的必備項目，並考慮系統的需求相互對照而制定的。依照CMARC4的規範，一筆書目記錄由多個欄位所構成，而其欄位是由欄號、指標、分欄三個部分組合而成。本研究依照CMARC4的特性及其資料結構，在設計DTD時有三個要點如下：

1. 指標為每一欄之第一個資料單元，用以指示該欄之內容。所以並不包含任何的資料，所以宣告為空標籤。

2. 為了使DTD可讀性增強，讓不熟悉CMARC的人也能了解，除分欄值使用英文字母外，其餘標籤皆用*CMARC4*中所定義的中文名稱。

註36 楊翔淳、林信成，〈以XML與CMARC簡篇為基礎之編目模組設計〉，圖書與資訊學刊(即將出刊)。

註37 魏令芳，〈簡略編目的發展與趨勢〉，大學圖書館，4：1，(民國89年3月)：頁114。

註38 陳昭珍，〈XML, Metadata與檔案資料數位化〉，可得自<http://archives.sinica.edu.tw/main/article06.html> (1 Mar. 2002).

註39 中國機讀編目格式修訂小組，中國機讀編目格式=*Chinese MARC Format* (附錄)，第4版 (國家圖書館，民國86年)，頁51-55。

3. 以*CMARC4*簡篇格式為主，再參考國圖NBINet核心書目記錄必備項目，交叉對照後，整理出DTD中必備項目(註40)。

此外，為了展現XML在結構化資料組織方面之優越性能，我們設計了一個編目與檢索系統，其目的在探討XML應用於圖書館自動化之潛力。此系統共分有六個模組：

1. 編目管理模組：管理者可新增、刪除、修改、查詢資料。
2. 資料查詢模組：提供使用者查詢資料。
3. XML轉換模組：將資料庫的資料轉換為XML格式。
4. DTD驗證模組：引用DTD驗證轉換所得之XML文件是否正確。
5. XML格式呈現模組：將DTD驗證通過之資料以XML格式呈現。
6. HTML格式呈現模組：將查詢所得資料以HTML格式顯示。

圖4 編目資料編輯畫面

本研究單元所提出之CMARC4簡篇DTD僅為實驗用途，並非公開討論後所制訂之共同標準，其目的除供本系統作為資料庫結構外，對於將來國內相關單位若欲制訂CMARC DTD之規範與標準，亦可作為一個參考用的草案。配合網路與資料交換的發展之下，CMARC DTD的制定是必然的，只是一個標準的訂定並不是由個人或少數人的力量即可完成，需經歷各相關單位的討論、協商、認可，且國內外各領域的發展也是考慮因素之一，如何讓CMARC DTD與其他的DTD做相關性的對照，或是彼此間有個參考、轉換的標準，有待各單位的互動及配合，未來的發展還有很長的一段路要走！

㈣ 圖書館行動線上公用目錄系統WAPOPAC(註41)

接下來的這一個研究單元中，我們從圖書館提供行動資訊服務的角度切入，

註40 *CMARC4*簡篇DTD之設計細節及詳細內容，請參見註36。

註41 林信成、楊翔淳，〈WAPOPAC系統設計與行動圖書館通訊技術之探討〉，圖書與資訊學刊，44期(民國92年2月)。

探討「無線應用協定」(Wireless Application Protocol，簡稱WAP)技術在圖書館之應用，並以WAP為基礎建構一個「行動線上公用目錄系統」(WAP-based OPAC，簡稱WAPOPAC)，再開發WAP中介軟體，搭配XML應用語言之一的「無線標示語言」(Wireless Markup Language，簡稱WML)，建構行動線上公用目錄網頁，將圖書館的線上公用目錄服務延伸至行動通訊網路上，讀者只要使用WAP終端設備，即可檢索圖書館的書目資料庫。

WAP延伸現有Internet上的標準並加以簡化，以適合手機的特性，利用行動通訊網路，以WML的語法，將資料傳送到手機等手持式配備(Handheld Device)上。WAP是一個全球通用的開放式應用層協定，其宗旨在於方便使用者透過行動通訊設備(如行動電話、呼叫器、雙向無線電.....等)存取網路資源。由於WAP是屬於應用層的協定，因此可架構於諸如GSM、GRPS、PHS、TDMA、FLEX.....等行動通訊系統之上(註42)。 WML則延伸自XML，可用來撰寫WAP網頁。由於WML是使用XML規格所制定出來的標示語言，所以WML文件亦滿足Well-Formed和Valid的條件。在WML文件一開始必須將XML宣告與文件格式宣告加入，另外在文件格式宣告當中，必須引用WAP Forum所制定的WML之DTD，以驗證WML文件是否符合規範。一份WML文件又可以分割成數張獨立的「卡片」(card)，WML網頁可以設定成多卡式，所有的卡片合起來統稱為卡片組(deck)，但WML瀏覽器一次只會顯示一張卡片的內容。換言之，一份WML文件包含一個卡片組，而在卡片組中包含一張或數張的卡片。WML文件之所以要以卡片來切割網頁內容，完全是因為行動終端設備螢幕無法一次瀏覽大量資訊(如手機、PDA等)，必須經過卡片的處理才得以將完整的資訊分批呈現在小螢幕上！再者，行動電話一般以輕薄短小為訴求，在硬體設備上如記憶體容量、螢幕大小、CPU運算能力、電源供應、輸入方式等方面有所限制(註43)。因此我們在WAPOPAC系統的設計上，無法如WebOPAC般複雜，僅能提供最簡便的查詢功能及顯示結果，因此本研究所建置的WAPOPAC系統(以下簡稱本系統)，僅提供三個主要查詢項目，分別是：

* 書名查詢
* 作者查詢
* ISBN查詢

本系統規劃的第一張WML卡為首頁的歡迎畫面，稱為「歡迎卡」(Welcome card)；第二張卡為查詢模式選單，稱為「選單卡」(Menu card)，可提供書名查詢(Title Search)、作者查詢(Author Search)、ISBN查詢(ISBN Search)共三種查詢選擇；第三張卡則視使用者所選擇的查詢模式而提供不同的輸入畫面，稱為

註42 林信成，網路概論與Internet實務應用(台北市：文魁，民國91年12月)，節6-6-4。

註43 姜景娟、陳尊明、林盈達，〈WAP行動上網技術分析與發展方向〉，網際網路技術學刊，2：1(民國90年1月)：頁41。

「查詢卡」(Search card)，又分「書名查詢卡」(Title search card)、「作者查詢卡」(Author search card)和「ISBN查詢卡」(ISBN search card)三種，其中，Title與Author查詢模式皆可輸入中英文供查詢，而ISBN查詢則須輸入10位數字的ISBN碼，以供系統比對；至於第四張卡之後則為查詢結果的輸出畫面，稱為「回應卡」(Response card)。

為了驗證系統運作是否順利，我們首先透過Nokia WAP Toolkit手機模擬器，實際測試系統運作。接著，我們以Motorola T191 WAP手機並搭配中華電信的門號，進行實機上網測試。在與前述的模擬器畫面比較後，可發現在不同的手機或模擬器中，所顯示的系統畫面並不完全相同。這是因為各廠家WAP設備對WML文件內容的解讀不同的緣故，如同各家Web瀏覽器對HTML文件中的標籤定義支援程度不同，以致同一份文件，在不同的瀏覽器上瀏覽時，顯示的結果會有出入！

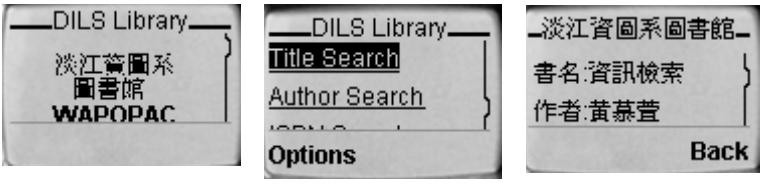

圖5 模擬器（Nokia WAP Toolkit）執行結果

圖6 實機（Motorola T191）測試結果

探討圖書館行動資訊服務的文獻，在國內外皆屬起步階段，而國外則較國內多些。例如，英國什羅浦郡立公共圖書館(註44)與漢普郡立公共圖書館(註45)，在其WAP服務中提供區域內的圖書館資料查詢，以WAP手機即可查詢各圖書館的地址、電話、開放時間等資訊；國內圖書資訊學界探討WAP技術應用於圖書館的文獻，至本文完成為止(民國92年2月)，有陳貞妃教授發表過以XML擷取資料、以WML表現資料的相關文章(註46)，以及關於無線應用通訊協定之行動圖書館架構

註44 什羅浦郡立公共圖書館WAP網頁，可得自<http://wappy.to/library>。

註45 漢普郡立公共圖書館WAP網頁，可得自<http://wap.hants.gov.uk/library/>。

註46 陳貞妃，〈行動通訊與資訊擷取之結合— 無線應用通訊協定與XML之探討〉，資訊傳播與圖書館學，7：3(民國90年3月)：頁104。

(註47)。但這些研究有的僅止於理論探討，有的僅提供單向資訊傳播，並未如本系統一般，提供有關圖書館館藏資訊的雙向互動式查詢。本研究透過系統實作，將OPAC與WAP相結合，利用原有的Web伺服器環境，建構了一個WAPOPAC實驗系統，讓使用者可經由WAP行動通訊設備，查詢圖書館的書目資料，進而利用圖書館資源。雖然其系統限制頗多，不如Web般的靈活與實用，但在行動通訊的蓬勃發展之下，圖書館對這方面的技術與應用應再進行更深入的探討，並透過各種系統實證的方式，才能更清楚其可行與不可行之處。20世紀末，網際網路和行動通訊已深遠的影響著人類溝通及互動的行為模式，這兩大領域的整合是否將再度影響圖書館的質變，引發讀者對行動資訊服務(Mobile Information Service)的需求，甚至帶動行動圖書館(Mobile Library)的發展，實在是一個頗值得更深入探討的課題。

五、結論與建議

本文首先從系統智慧化與文件智慧化兩個領域逐漸匯流的角度切入，探討電子文件的智慧化程度實是影響檢索系統性能的重要因素；接著，從XML發展趨勢來看，我們可以很清楚的發現，以XML為核心的技術已經扮演了提升文件智能的重要角色；再者，藉由一系列的實作，我們以XML為核心，首先在網路上建置了兩家虛擬書店，彼此透過XML進行資料的傳遞與交換，驗證了XML在Web資料交換上的便利，若將此系統加以擴充，則可成為N個彼此透過XML進行資源共享的一般化整合系統；再透過新聞管理與出版系統的實作，將新聞內容以XML作結構化的處理，並自訂Metadata來描述其語意，搭配資料檢索模組，可確實針對新聞內容作精確檢索，其精確度優於傳統的全文檢索結果；再以中國機讀編目格式簡篇第四版為標準，制定其適用之DTD，透過系統實作，架構一個基於XML的簡易編目系統，對於小型及地方圖書館來說，不但足以滿足讀者的需求，且小型自動化系統容易開發，可省下相當龐大的經費；最後將OPAC與WAP/WML相結合，建構了一個WAPOPAC行動線上公用目錄系統，讓使用者可經由行動通訊查詢圖書館的書目資料，隨時隨地利用圖書館資源。

有史以來，圖書館的運作便與資訊科技的發展存在著密不可分的連動關係，從人工作業到自動化處理；從紙本資料到電子資源再到智慧型電子文件；從卡片目錄到線上公用目錄(Online Public Access Catalog，簡稱OPAC)到WebOPAC，再到本研究所提出的WAPOPAC；無一不展現出圖書館求新求變的本質，也揭露了圖書館乃一珍惜過去、立足現在、放眼未來，融合了傳統與先進、文化與科技的有機體。

註47 陳貞妃，〈基於無線應用通訊協定之行動圖書館架構〉，資訊傳播與圖書館學，8：4(民國91年6月)：頁40-41。

總之，本研究且在實際個別建構系統之後，更能深入瞭解以XML為本的智慧型文件之特點，及其與智慧型系統整合之優勢，並且藉由系統實作的驗證，更加深應用XML的信心，期望在不久的將來，XML發展更成熟之際，能全面的應用於網路之上！

誌　謝

本研究得以順利進行，承蒙國家科學委員會經費補助(計畫編號：NSC 90-2413-H-032-011)，研究助理楊翔淳、陳勇任全力協助，特此感謝。

A Research on the Integration of Intelligent Document and Intelligent System

Sinn-Cheng Lin

Associate Professor
Department of Information and Library Science, Tamkang University
Taipei, Taiwan, R.O.C.
E-Mail : sclin@mail.tku.edu.tw

Abstract

This study focuses on the integration of intelligent documents and intelligent systems. First, the paper defines the intelligent document as an electronic document that has extra self-description information, semantically. We believe that the intelligence of the document would be an important factor that impacts the performance of information retrieval systems. Next, by exploring the development of XML, we find that the XML-based technologies already became the principle of intelligent documents. Moreover, a series of system implementations have been done in this paper, they are a data exchange system, a news publication system, an XML-based CMARC cataloging system and a WAP-based OPAC system. These experiments demonstrate the application potentials of XML in many fields, such as electronic data exchange, electronic publication, library automation and mobile information service of library.

Keywords : *XML; Intelligent publication; Intelligent document; Intelligent system; WAPOPAC*