支援助理型軟體之電子圖書館（III）

Digital Library for Information Agents

計畫編號：NSC88-2213-E-002-006

執行期限：87 年 8 月 1 日至 88 年 7 月 31 日

主持人：陳文進　台灣大學資訊工程系教授

wcchen@cmlab.csie.ntu.edu.tw

一、中文摘要

Internet 提供了大量垂手可得的資源，但也令使用者難以找尋所需資料．目前在 Internet 上雖然有各式各樣的資源搜尋工具，但它們都無法提供使用者一致的分類空間，並會造成網路的沉重負荷．數位圖書館是將數位資料分類處理，並提供各種工具，以利資訊交流與搜尋的環境與架構．在這個架構下，資訊可以快速而有效的流通，進而加速技術的發展與產業的升級，可說是國家資訊基礎建設的關鍵技術．

本計劃為『支援助理型軟體之電子圖書館』三年計劃之第三年計劃， 其目標在設計並實作一個數位圖書館的架構，以支援助理型軟體，成為資訊擷取助理的骨幹，使其能透過一致的界面，擷取各類的網路資源．在上一年度中， 根據這個構想， 我們完成了一個虛擬圖書館， 其不僅為我們計劃提供了雛型架構， 更提供了一個強有力的搜尋引擎．

而未來我們整個電子圖書館的架構是以多個分散式的代理者為主，整合智慧型助理及外界的資料庫，構成一完整的環境．代理者由查詢管理員、書目資料伺服器、內容註冊處及路由器等四大模組組成，是我們研究的重點．使用者可透過智慧型助理，以圖形使用者界面與互動模式和代理者溝通，以獲得所需之資訊．作者或圖書館員則可透過書目資料伺服器輸入高品質的書目資料，或將資料直接輸入其專屬的資料庫系統中．無論對使用者或作者而言，都只見到單一的界面與一的資訊分類，無須考量資料所在的位置，及不同的查詢方式和通訊協定．我們認為此架構將具有易於使用、可擴充、高效能等性質．

有關資訊擷取助理部份請見整合型計劃『助理型軟體系統之研製與應用』．

二、英文摘要

Internet has lots of resources， which makes users hard to find the desired information. There are many resource discovery tools trying to solve this problem. Most of them do not support cataloging of information and can put heavy burden on the network. Digital Library is an architecture and environment that provides various tools to facilitate searching and exchanging of information. Information can be exchanged fast and efficiently in this environment， such that techniques can be evolved and industry can be promoted. Digital Library is also an important component of National Information Infrastructure.

This project is trying to design and implement a framework of digital library, which provide an uniform interface for information retrieval agents as their backend to access various kinds of Internet resources. The framework will consist of many distributed brokers， intelligent agents and external databases. The broker itself includes four major modules： query manager， bibliography records server， content registry and content

router. It is our major research area. Users can obtain the desired information by communicating interactively with brokers through intelligent intelligent agents and graphic user interfaces. Authors and librarians can publish high quality bibliography data to a broker, or just insert bibliography data to their private databases. In this architecture, users and publishers can access various kinds of resources in an uniform information space and by an unique interface without knowing the locations, the query languages and the access protocols of the desired data. We believe the architecture would have the properties of usability, scalability and efficiency.

Please see "The Implementation and Application of Agent Software Systems" for more information about information retrieval agents.

三、計畫緣由與目的

本計劃為『支援肋理型軟體之電子圖書館』三年計劃之第三年計劃. 電子圖書館的構想來自近年來 Internet 蓬勃的發展. 爾來 Internet 不論在使用人數或資料數量上都以驚人的速度成長, 其應用領域也從學術擴展到商業及日常生活之中. 這樣的資訊革命將以徹底改變我們未來的生活形態, 使得電腦, 通訊與消費性電子產品(Computer, Communication & Consumer electronics, 3C)結合在一起, 成為日常生活的一部份. 在這個全數位化的環境裡, 傳統的商業行為和生活型態將受到嚴厲的挑戰. 消費者將從網路上擷取到所需的資訊, 改變目前大眾傳播主導商業行為的現象. 這樣的小眾傳播, 將促成個人化商品的出現, 使得小廠商無需投注大量廣告經告經費也能和大廠商競爭. 我們可以預見許多小型甚至個人企業的興起, 郵購商品的拓展, 以及傳統中間商的沒落.

台灣缺乏自然資源, 經濟發展有賴進出口貿易, 因此中小企業的彈性和活力成為經濟發展的主力. 但在資訊工業蓬勃發展的今日, 如何掌握資訊技術, 善用網路資源, 開發新技術與產品, 便成為我們將來繼續生存發展的關鍵. 我們首先的工作就是將資料數位化, 並提供良好的架構與搜尋工具, 使資訊能快速的流通, 進一步轉換為有用的佑識, 以加快新產品開發的速度, 提供消費者更好的服務, 藉以創造新的商機.

數位圖書館是將數位資料分類處理, 並提供各種界面一致的搜尋工具, 以利資訊交流的環境與架構. 它同時也是國家資訊基礎建設(National Information Infrastructure, NII)的關鍵技術. 因此近年來 國內外相關的研究便蓬勃發展, 各大軟體公司更積極投入, 以期在網路市場上佔有一席之地.

目前 Internet 上已有許多的搜尋工具, 如已商業化的 archie, yahoo, infoseek, 以及其它學術研究的成果, 如 WAIS, lycos, harvest, WWWW, WebCrawler 等等. 然而這些工具都有其設計上的限制與缺點, 我們可以歸納如下:

1. 缺乏分類:我們所要搜尋的資料通常具有一定的相關性, 這也是傳統圖書館分類的依據. 透過這樣的分類, 使用者可以很容易的找到未佑但相關的資訊. 目前的搜尋工具大多針對 ftp 或 WWW 伺服器的文件做關鍵字索引, 並未提供高品質的書目資

料， 因此在資訊的過濾及相關訊的找尋上無法達到傳統圖書館的水準.

2. 無法提供全面的資料：目前的工具大多採取定時詢問 WWW 伺服器的方式來建立一集中式的索引， 如此一來便無法趕上 Internet 資料的成長速度， 造成各種工具只能見到部分的網路資訊.
3. 缺乏擴充性：任何集中式的工具都無法容納不斷增加的資料量和使用者. 而 Internet 呈現幾何倍數成長的使用者和資料量， 將迫使所有的系統採用分散處理的技術.
4. 增加網路的負擔： 利用定期詢問的方式來建立集中式索引的工具， 會造成網路不必要的負荷， 因為索引比文件本身小得多， 且定期詢問會造成重覆處理已處理過的文件. 若有多種同性質的工具同時存在， 將使問題更加嚴重.
5. 無法與其它的資料庫系統相溝通： Internet 的資源並不是只有 ftp 和 WWW 而已. 它還包括了圖書館的書目資料庫， 各種商業線上資料庫，政府部門資料庫及各種科學資料庫等. 理想的電子圖書館應能整合各類的資料庫, 以發揮最大的效益.
6. 缺乏智慧：無法根據使用者的文化背景和習慣做適當的調整, 以符合使用者的需求.

由於這些問題的存在以及解決這些問題的迫切性, 國外有許多研究計劃正在進行. 其中最重要的有美國國家科學基金會自 1994 起資助 UC Berkeley, Stanford University, U. of Illinois, U. of Michigan, UCSB 與 CMU 等學校的數位圖書館先導性計劃, 英國大英圖書館自 1993 年起之先導性存取計劃, 以及美國國會圖書館自 1990 年起之電子圖書館計劃.

本子計劃所在之整合型計劃『助理型軟體系統之研製與應用』, 主要是在發展數個助理型軟體系統. 其中最重要的是建構一個具有智慧型助理功能的電子圖書館, 以便使用者能找到真正需要的資料, 正確的使用資料, 進一步將資料變成有用的知識. 因此本計劃較上述各研究中的電子圖書館計劃有更強大的能力, 也因此需要探討更多的理論與系統架構.

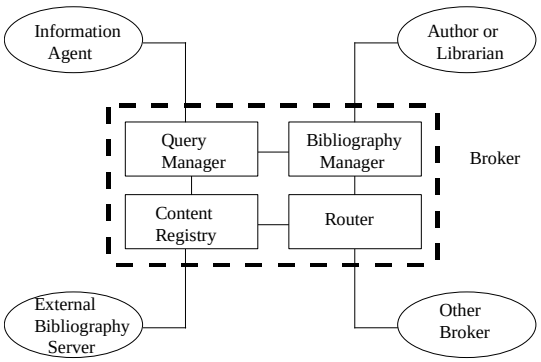
四、計畫實行步驟

本計劃為『支援肋理型軟體之電子書館』三年計劃之第三年計劃．計劃中的電子圖書館即下文所稱數位圖書館， 我們認為其應具備下列幾點特質， 發揮所謂數位的功能，方能稱為一個好的數位圖書館：

1. 易於使用. 這是所有成功系統的必要條件. 在資訊搜尋的領域裡， 這意味著它可允許使用者以互助的模式和系統溝通， 並將資訊所在的位置及擷取通訊協定隱藏起來， 讓使用者以其熟悉的分類方式來尋找所需的資訊. 同時它也要能依據使用者的習慣和文化背景， 做適當的調整.
2. 可擴充. 這個架構必須能理不斷增加的使用者人數及資料量， 並能隨時加入所需的計算能力， 同時能因應不同組織的管理與安全要求.
3. 高效能. 此架構必須能快速反應使用者的查詢， 並降低對網路頻寬的需求， 不因使用者人數及資料量的增加而使反應變慢， 或造成網路的擁塞.
4. 能整合其它資料庫系統. 此架構應具有足夠的彈性以整合各種資料庫系統之通

訊協定和資料格式，以因應快速變遷的網路環境.

針對這些需求，我們將以圖一的架構來完成.



圖一 數位圖書館系統架構圖

圖一的架構是以代理者(Broker)為主，其內包含了查詢管理員(Query Manager)，書目資料伺服器(Bibliography Server)，內容註冊處(Content Registry)及路由器(Router)等四大模組. 代理者可以彼此交流，並和外界的書目資料庫(External Bibliography)共同構成一個電子圖書館系統. 使用者可透過智慧型資訊擷取助理，以圖形使用者界面與互動的模式和數位圖書館溝通. 文件的作者或圖書館員則可透過書目資料伺服器輸入高品質的書目資料，或將資料直接輸入其專屬的資料庫系統中. 無論對使用者或作者而言，均只見到單一的界面，無須考量資料所在的位置，及不同的查詢方式和通訊協定. 代理者各模組的功能和其間流程敘述如下：

1. 查詢管理員：查詢管理員負責處理智慧型助理所送來的查詢命令，將該命令和內容註冊處相比較，以找出可能會有所需資料的書目資料伺服器，然後將查詢命令依該伺服器所接受的協定做適當的轉換，再交由該伺服器處理. 然後所得的結果，轉換成一致的格式，再傳回給智慧型助理.
2. 書目資料伺服器：書目資料伺服器管理作者所輸入的書目資料. 它必須能同時服務多個使用者，並能以分散式處理的方式管理大量資料. 同時它必須能處理不同的書目資料格式，如 MARC，TEI，URC 等等.
3. 內容註冊處：此模組管理外界書目資料庫的宣告資料. 這些宣告包括擷取協定，如 Z39.50，CGI 和 ODBC; 所含內容之分類，如主題，地區和時間等; 以及欄位名稱的對照關係等. 當查詢管理員收到查詢命令時，必須先參考此模組以找出相關的伺服器，再依其存取協定做適當的處理.
4. 路由器：路由器最主要的功能就是將內容註錄複製到所有的代理者. 雖然我們只複製了伺服器的內容宣告，並未複製其資料，但仍需有良好的複製架構，以確保系統的可擴充性，並進一步降底網路的負荷. 同時此複製架構必須能彈性增減代理者，並處理網路臨時或永久性的斷線，以確保系統的運作及一致性.

為了實現上述的架構設計，我們在 Unix 環境下研究並實作下列的工作：

1. 內容註冊處之格式訂定及擷取協定之實作. 主題宣告部份採用美國國會圖書館分類法，擷取協定將實作 CGI 及 Z39.50.
2. 路由器架構與傳輸協定之研究. 此部份以軟體模擬的方式來驗證其效能與可行性.
3. 查詢命令之訂定，與如何找尋相關伺服器之研究.
4. 書目資料伺服器之製作. 可輸入 MARC，URC 格式之資料，並能同時

服務多個使用者

五、本計畫成果:

建構一完整的數位圖書館架構. 在此架構下, 數位資訊可以有效的流通, 以促進學術的發展, 進一步帶動產業升級. 未來它也可以做為商業廣告的媒介, 增進我國介業的競爭能力. 此外參與本計畫的人員, 除了對各種擷取協定及書目格式有深入的瞭解外, 對資訊擷取, 分散式系統, 網路架構, 系統實作與整合等相關知識, 將有全盤完整的掌握.

我們第三年完成之工作如下:

1. 實作路由器模組，並確保系統的擴充性，可彈性增減代理者，以處理網路臨時或永久性的斷線
2. 整合代理人各模組，並完成功能測試
3. 與智慧型助理及外界資料庫之整合
4. 數位資料庫及實驗平台之擴充, 與系統效能之調整
5. 數位圖書館架構之推展及相關應用之研究

六、參考文獻

1. Ron Daniel, Terry Allen, “An SGML-based URC Service”. IETF draft, draft-ietf-uri-urc-sgml-00.txt.
2. William Arms, David Elly, “The Handle System”. IETF draft, draft-ietr-uri-urn-handles-00.txt.
3. Ron Daniel, Michael Mealling, “URC Scenarios and Requirements”. IETF draft, draft-ietf-uri-urc-req-01.txt
4. Paul E. Hoffman, Ron Daniel, “Trivial URC Syntax: urc0”. IETF draft, draft-ietf-uri-urc-trivial-00.txt.
5. Darren R. Hardy, Michael F. Schuwartz, Duane Wessels. “Harvest User‘s Manual”. Technical Report CU-CS-743-94, University of Colorado at Boulder.
6. Darren R. Hardy, Michael F. Schwartz, “Essence: A Resource Discovery System Based on Semantic File Indexing”, 1993 Winter USENIX, San Diego, CA.
7. Peter B. Danzig, Dante DeLucia, Katia Obraczka, “Massively Replicating Services in WideArea Internetworks”. Available from ftp://catarina.usc.edu/pub/kobraczk/flood.ps.Z
8. James D. Guyton, Michael F. Scuwartz, “Location Nearby Copies of Replicated Internet Servers”. Technical Report CI-CS-762-95, University of Colorado.
9. C. Mic Browman, Peter B. Danzig, Udi Manber. Michael F. Schwartz. “Scalable Internet Resource Discovery: Research Problem and Approaches”. Communications of the ACM., August 1994, Vol.37, No.8, pp98-107.
10. Lorcan Dempsey. “Network Resource Discovery: A European Library Perspective”. Libraried, networks and Europe: a European networking study: Neil Smith(ed). London: British Library Research & Development, 1994.
11. William P. Birmingham, Edmund H. Durfee, Tracy Mullen, and Michael P. Wellman. “The Distributed Agent

Architecture of The University of Michigan Digital Library". DL-lib, July, 1995.

12. Bipin C. Desai. "The Semantic Header and Indexing and Searching on the Internet". http://www.cs.concordia.ca/~bcdesai/cindi-system-1.0.html
13. Diane Vizine-Goetz, Jean Godby, Mark Bendig. "Spectrum: A web-based Tool for Describing Electronic Resources." Computer Networks and ISDN Systems 27:985-1001.
14. Joann Janet Ordille. "Descriptive Name Services For Large Internet." Ph.D thesis, University of Wisconsin-Madison, 1993.
15. Martin Roscheien, Christian Mogensen, Terry Winograd, "A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples". Technical Report CSDTR/DLTR, Stanford University.
16. Henry M. Gladney etc. "Digital Library: Gross Structure and Requirements". IEEE Computer Society Press, Proc. Workshop on On-line Access to Digital Libraries, 1994.
17. Michael F. Schwartz etc. "A Comparison of Internet Resource Discovery Approaches". Computing Systems. 1992.
18. Ray R. Larson. "Design and Development of a Network-Based Electronic Library". Navigating the Networks: Proceedings of the ASIS Mis-Year Meeting Portland, Oregon, May 21-25, 1994.
19. Edward A. Fox. "Users, User Interfaces, and Objects: Envision, a Digital Library". Journal of the American Society for Information Science, Sep. 1993, pp480-491.
20. "Digital Libraries". Communications of the ACM, April 1995, vol 38, num. 4.

1. a a