數位圖書館中自動知識擷取之研究

Study of Automatic Knowledge Extraction in Digital Libraries

計畫編號：NSC89-2213-E-002-019

執行期限：自 88 年 8 月 1 日至 89 年 7 月 31 日

主持人：歐陽彥正　　台灣大學資訊工程學系　　教授

Email: yjoyang@csie.ntu.edu.tw　　Fax: (02)2368-8675

中文摘要：

本計畫的研究重點為如何自動擷取涵藏於數位圖書館內容的知識，以發揮數位圖書館在知識管理與應用上的積極意義。提出本研究主題的動機主要源自於下列觀察：數位圖書館除了提供基本的內容儲存與檢索的功能外，更積極的意義是能藉助 knowledge discovery 的方法，幫助人們發掘涵藏於數位圖書館內容中的知識。這些 knowledge discovery 的方法能大幅增加人類對知識管理及應用的能力，同時也是數位圖書館對資訊化社會所能發揮的最積極的功能之一。本研究將以收集在「台大數位圖書館/博物館」中的歷史資料作為知識擷取的標的。所擷取的知識則將記錄在我們為歷史性知識所建立的 object model 中，以作為進一步自動知識推導的基礎。以上所討論的自動知識擷取及推導的機制能有效提升人們整理甚至發現新知識的能力。這些自動知識擷取及推導的機制雖然就人的智慧能力觀之，可能是相對簡單的。然而由於數位圖書館的資料量相當大，而所隱含的知識量亦相當可觀，因此要有系統的整理出這些知識，所需耗費的人力必然相當可觀，也相當不經濟。本計畫的目的即是以歷史性資料為範本，希望證明知識自動擷取與自動推導機制結合數位圖書館在未來資訊社會中知識管理與應用方面所能發揮的重大功效。

在研究成果方面，我們已經可以相當成功地自表列式文件中，自動擷取有興趣的知識，如一個人的字、號、籍貫等，以獲得初步成果。經自動知識擷取機制所獲得的知識，我們將再應用自動知識推導過程，建構出物件間的關連網路。

英文摘要：

The main objective of this project is to study automatic knowledge extraction based on digital library contents. The significance of this project is that its results can greatly enhance the applications of digital libraries in knowledge management. This project was motivated by observing recent development in knowledge discovery research. This project will use the contents in the National Taiwan University Digital Library and Museum (NTUDLM) as the target. The knowledge extracted will first stored in an object model that we have developed. Then, an AI production system will be applied to explore and derive embedded knowledge. To human being, the knowledge extraction and deduction mechanisms studied in this project may be quite straightforward. However, if the digital library contains a large

quantity of contents, then this task will become very exhaustive to human being. One of the main goals of this project is to demonstrate how knowledge extraction and deduction mechanisms can help and how they can enhance the applications of digital libraries in knowledge management.

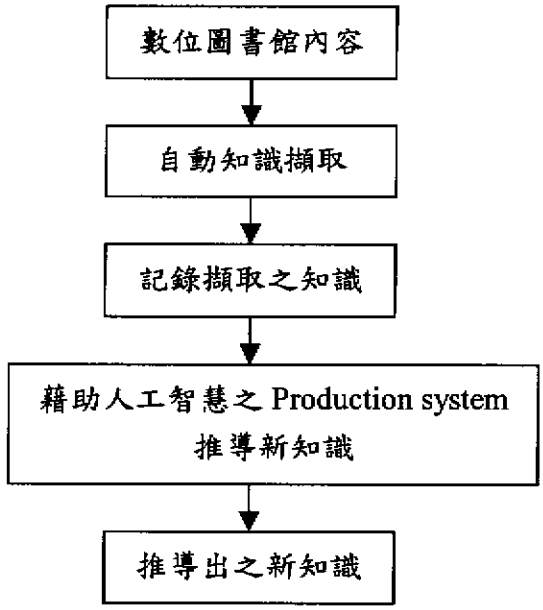
In this project, knowledge extraction issues for digital library content were investigated. The study was conducted along with implementing an prototype system. The prototype system provides a platform for evaluation some aspects of knowledge extraction which will provide insights for further improvement.

計畫緣由與目的：

觀察電腦科技約五十年的發展，可以歸納出下列幾個階段。在第一階段，以大型主機為核心的電腦科技主要被應用在科學、工程、軍事，與特定商業的計算上。這個時期，只有電腦專業人士才具備操作與使用電腦的技能。到了第二階段，個人電腦與工作站興起後，一般人士才逐步借用電腦完成部份工作。但在這個階段，電腦對一般人而言仍只是輔助性的工具，而非生活或工作的重心。時至今日，由於網際網路的興起，創造了新一代的通訊方式。電腦科技提供的不再僅限於計算與資料的選取，進一步的，電腦已成人們獲取資訊的重要管道，而人類社會也將進入全面的資訊化。

由於電腦成為文明社會中資訊獲取的重要管道，『資訊提供站』的建構便成為這整個發展步驟中的關鍵工作。這也是為何近年數位圖書館/博物館的建構已成為各方面人士積極推動之工作的主要原因。

數位圖書館/博物館所發揮的功能除了『資訊提供站』外，更積極的是它可以藉由 knowledge discovery 的方法，大幅增進人類在知識管理與應用上的效率及能力。然而要做到 knowledge discovery，首先必須能夠擷取涵藏在數位圖書館的知識，同時以一個適當的 model 將所發掘的知識記錄下來，再應用自動推導的機制發現及整理新知識。下圖所示即為整個 knowledge discovery 流程。



上圖所談到的自動知識推導機制在人工智慧領域中已研究多年並有相當成果，因此將不在本計畫的研究範圍。至於記錄知識的 model，本研究團隊在過去一年間已建構出一個基本架構，這部份的結果將在稍後討論。而本計畫的研究重點則在知識自動擷取機制方面。

本研究團隊為歷史性資訊及知識發展出的 object model 將用以記錄由自動知識擷取機制所發掘整理出的知識。而後再由人工智慧的 production system 進一步推導出新知識。

有關資訊/知識擷取機制的研究是近年來資訊領域中相當受到重視的一個課題。由於發展一個通用性的資訊/知識擷取機制有相當高的困難度，因此大多數的研究均是針對特定範圍而作的。本計畫的目的是以「台大數位圖書館/博物館」的內容為基礎，期望發展針對歷史文獻的知識擷取機制，並配合人工智慧的知識推導機制，使數位圖書館進一步發揮在知識管理與應用上的積極角色。

研究方法與成果：

如前所述，欲發展一個通用性的資訊/知識擷取機制，有相當的困難度。因此一般的做法都是先選定特定的標的。本計畫將以台灣方志為主要標的。由台灣方志中記載有關清代在台灣任職的官員簡歷表中可以觀察到一些簡單的 pattern。根據這些 pattern 便可藉助程式自動擷取出每位官員的籍貫，到任時間，離職時間。本計畫由較簡單的表列式記載開始，研究自動知識擷取的機制，再逐步擴展到以敘述式記載為擷取標的。目前針對表列式記載所發展的自動擷取機制相當成功，因為 pattern 相當明顯且固定。而針對敘述式記載所發展的自動擷取機制則只能針對範圍相當明確的資訊（如一個人物的字、號、籍貫）做出部份成果，且其精確性及完整性尚待進一步檢驗。

台灣縣知縣	錢業南河南彬州人。康熙三十一年任；三十三年以劾去。 李中素湖廣麻城人，由貢生。康熙三十四年任；卒於官。 盧承德號克明，鑲黃旗人，由官生。康熙三十六年任。延平府沙縣調補。 陳璸號眉川，廣東海康人，由進士。康熙四十一年任；秩未滿，陞河南新鄭縣知縣。
縣丞	蔣以仂號千選，紹興府山陰縣人，由功貢。康熙三十五年任。福州府閩縣調補。 章祖祺浙江山陰縣人，由監生。康熙四十四年任；秩滿，陞江南歙縣縣丞。 汪立忠號恕齋，建寧府甌寧縣人，由廣。康熙四十四年任；調補湖廣武陵縣。
典史	婁克仁浙江紹興府會稽縣人，吏員，延平府龍溪縣。康熙三十五年調補。 孫日昇山東莒州人，由吏員。泉州府德化縣典史調補。康熙四十年任。 李廷貴號閩嘉，江南桐城人。由監生。漳州府龍溪縣典史調補。康熙四十四年任。 陳茂文號煥菴，湖廣江陵籍，浙江餘杭人，由吏員。寧德

表一

至於具體研究成果，我們已經可以相當成功地自表列式文件（表一）中，自動擷取有興趣的知識，如一個人的字、號、籍貫等等（表二）。經自動知識擷取機制所獲得的知識，我們將再應用自動知識推導過程，建構出物件間的關連網路。如前所述，這些應用自動知識擷取及推導機制所獲得的知識，原先可能散落及隱藏在大量文獻中的各個角落。因此要運用人力

去逐一發掘將相當沒有效率且極不經濟。我們並不預期，這些自動知識擷取及推導機制將會完全取代人的思考推理，我們只是認為人應更善用電腦所不能及的思考推理能力去發掘及整理更高層的知識。這些對人而言，相對簡單但繁瑣的知識整理及思考推理工作交由電腦處理即可。這樣人才能更專注於高層的心智活動。

姓名	職稱	號	籍貫
錢巍業	台灣縣知縣		河南彬州
李中素	台灣縣知縣		湖廣麻城
盧承德	台灣縣知縣	克明	鑲黃旗
陳璸	台灣縣知縣	眉川	廣東海康
蔣以選	台灣縣縣丞	千仞	紹興府山陰縣
章祖祺	台灣縣縣丞		浙江山陰縣
汪立忠	台灣縣縣丞	恕齋	江南歙縣
婁克仁	台灣縣典史		浙江紹興府會稽縣
孫日昇	台灣縣典史		山東莒州
李廷貴	台灣縣典史	閩嘉	江南桐城
陳茂文	台灣縣典史	煥菴	浙江餘杭

表二