行政院國家科學委員會專題研究計畫成果報告

全球資訊網資料發掘之研究

The Data Mining in a WWW Environment

計畫編號：NSC-88-2416-H-004-036

執行期限：87 年 8 月 1 日至 88 年 7 月 31 日

主持人：楊亨利教授　國立政治大學資訊管理系

一、摘要

近幾年來，全球資訊網在網際網路上蓬勃發展，因此而產生在網際網路上尋找資訊的問題。以目前的搜尋引擎與目錄服務的作法並無法找到真正有用的、潛在的資訊與知識。而有待運用資料發掘的技術於其上。但是，構成全球資訊網網頁之超文字遠比傳統的關連式資料庫非結構化。本研究先對資料發掘，全球資訊網資料搜尋等文獻與現況進行整理分析，並採用 OCLC **與** NCSA **所建議** Dublin Core **或** Desai（1997）**之** Semantic Header **為基本之超文字文件結構，再針對選定之應用領域（英語教育）建議彈性擴充之屬性。採用** Han（1995）**之概念樹及多層次資料庫的觀念，提出一可行之全球資訊網上資料發掘的整體架構；並發展一包含行業領域規則之資料發掘系統雛形，以驗証其可行性，並評估其日後實務操作之複雜性。**

關鍵詞：全球資訊網；資料發掘；知識發現；超文字文件結構

Abstract

In recent years, a number of applications on World Wide Web (WWW) by Internet have grown rapidly. A unresolved problem of finding information on Internet has emerged . It is difficult for current search engines or directory services to identify valid, novel potentially useful patterns in data. Someone might hope to apply the techniques of data mining (or called knowledge discovery). However, the web pages written in HTML (Hyper Text Markup Language) are more unstructured than traditional relational database. This research began with literature review of data mining and WWW data searching and also evaluated the current business practice. The Dublin Core proposed by American OCLC and NCSA (1996), or semantic header proposed by Desai (1997) was modified as our HTML basic structure elements. This research also suggested some augmented attributes in a selected application domain (the English education). The concept hierarchy and multiple layered database suggested by Han (1995) will be adopted. A feasible integrated architecture of industry domain rule was then developed to test the feasibility and evaluate the complexion of this kind of system in the future real world.

Keywords：World Wide Web, Data Mining, Knowledge Discovery, HTML Structure.

二、緣由與目的

資料發掘（Data Mining**）或稱資料庫中的知識發現** (Knowledge Discovery in DataBases)**乃資料庫中選擇合適資料、資料處理、資料轉換，資料發掘至結果評估，以獲得非顯然的發掘隱含的，前所未知的而可能有用資訊的過程【**8**】。而各種針對現行資料庫的資料發掘方法與技巧也紛紛提出，則廣泛地為人所討論【**8**】【**9**】【**10**】【**19**】【**21**】【**22**】。**Han **的概念樹學習法，便是近年來在資料發掘的**

研究中頗受重視的方法之一；其主要的精神所在，是由領域專家依其對於領域的了解，將領域知識存於概念樹中。而於資料發掘過程，將存在於資料庫的資料屬性，其中可以抽象化的部份依此概念樹不斷向上抽象化，直到抽象化之屬性個數符合預先設定之門檻值為止，再整理成邏輯規則。讓使用者所看到的，不是原始資料庫中零碎的資料，而是結合此領域的概念與術語而成的整體知識。目前關於概念樹學習法的探討，已經十分豐富，不但已經能解決由關連式資料庫中發掘出特性規則（列出某特定族群如研究生的特性）【11】、區別規則（比較多類族群如研究生與大學生的差異）【1】【2】、關聯規則（指出事物間的關聯性，如買牛奶與買麵包的關聯性）【3】【13】、分群規則（將資料庫依其特性分群）【5】、進化規則（表示資料在持續時間記錄下之變化趨勢）【12】等的問題。同時也已經對於將概念樹應用於地理空間【14】、物件資料庫【15】有過許多討論。

但隨著這幾年全球資訊網(WWW)在網際網路上的蓬勃發展，主要以超文字格式 (HTML)分散存在於網際網路上各公開或私有網站中的資料，也逐漸累積成為一個內容豐富而不可忽視的資料來源，因而成為找尋知識與資訊的重要目標之一。將全球資訊網與其上之超文字資料在與傳統資料庫與其內容資料兩相比較，顯然前者又較後者增加了（1）構成資料主體的超文字格式遠比傳統關連式資料庫的表格來的非結構化、（2）資料的分布遠較一般資料庫來的分散、（3）其資料內容所描述的領域可能比傳統資料庫中的資料領域更為廣闊與複雜、（4）其使用者也較傳統資料庫的使用者成長更為迅速且更難掌握其意圖等四項特性【16】。前兩項特性增加了從資料中獲取資訊的困難度，而後兩項特性則使知識之發掘變之更加不易。因此如何克服此四項特性，針對個別需求，而在其中找到符合需求的知識與資訊可以說是一個更具挑戰性的課題。我們可以將其稱之為全球資訊網上的資料發掘(Data Mining on WWW)或全球資訊網之資料發掘（Web Mining)）【7】。也有一部份的學者則將此一議題與過去舊有的網際網路資源上的資料與知識的尋找的研究相結合，稱之為網際網路的資源發現（Internet Resource Discovery）【4】【17】。為了輔助搜尋引擎與目錄服務發揮應有功能與消除其衍生障礙，後續的一些研究提出了一些方法加以改善。其一為 Han、Zaine 與 Fu【16】希望個別超文字文件之提供者能遵循某些資料結構提供資料，而後進行抽象化之萃取，並將其結果儲存為其個別之結構化之關聯式資料庫，再對這些個別資料庫進行收集與再抽象化之萃取；逐步構形成一多層狀之資料庫。當使用者希望由全球資訊網上的資料中找出資訊時，可以分別就其需要，直接對上層資料庫做相關的查尋，這樣就可以既快速又省事的得到所需的資料。如果一個這種全球資訊網的多層資料庫能成功建立，那麼就可以為全球資訊網提供一個層級組織井然的全域性的觀點，讓使用者不必直接從漫無組織的全球資訊網去擷取資訊。Zaine 將這樣的觀點稱之為一個虛擬的全球資訊網（VWV），並根據此一架構

與 Han 的概念樹演算法與資料擷取查詢語言(Data Mining Query Language, DMQL)，發展全球資訊網資料擷取語言（WebML）【20】。但對發掘之資料究竟要如何遵循哪種資料結構，他們並未提供滿意的答案。

針對以上的困境，有人主張在超文字格式的資料架構中加強其結構化之段落定義，使之具有網際網路資源描述與著錄意義內涵。企圖透過新的定義，使這些或由資料原作者填寫，或利用對非結構化本文分析之技術所產生而的結構化段落，以解決超文字格式資料之非結構化的問題。其中，較出名的相關研究包括由 OCLC 與 NCSA 整理出十二項具有網際網路資源意義內涵的資料項目，簡稱為「Dublin Core」【18】。Desai【6】則認為 Dublin Core 過於簡略與遷就現行架構，提出直接修正超文本語言之架構，徹底改寫超文字語言中較具結構性的 HEAD 標記段落，重新定義其子段落組成，而成為具有網際網路資源描述與著錄意義內涵的語意標頭（Semantic Header）。

值得注意的是不管 Dublin Core 或 Semantic Herder 不僅在全球資訊網上的資料中找出資訊的問題有所幫助，其所定義的網際網路資源意義內涵的資料項目也隱含一些網際網路與出版之知識，而具有資料發掘之價值。以 Dublin Core 為例，我們便可能可以從其十二個資料項目中找到諸如在特定日期下，某一個地區在一固定使用語言下傾向於撰寫哪一類網頁的特性規則、不同的出版者針同一文件涵蓋地區分別傾向於運用哪一種語言發表網頁之區別規則、文件涵蓋地區與發表網頁使用語言之間的關連規則、同一地區網頁之分群規則與隨出版日期同一涵蓋地區使用發表語言隨之改變的進化規則等具有意義的知識。

因此不管採用 Dublin Core 或 Desai 之 Semantic Header 均有助於解決 Han 與 Zaine 之最原始資料結構形成的問題，提供有用的資訊。這也是本研究之緣起，希望發展一個可行的全球資訊網上的資料發掘架構。當然，這裡基本的假設是超文字文件提供者願意填註 Dublin Core 或 Desai 之 Semantic Header。對於這點其實不必太悲觀，以 WWW 的發展，某種超文字結構標準的形成是必然的趨勢。而且在未來可延伸式標示語言（XML）的技術與標準更加成熟的環境下，Dublin Core 或 Desai 之 Semantic Header 可以更容易用直接定義的標記方式，儲存在使用者的網頁中。另外由於全球資訊網上資料領域廣闊與複雜，固定項目的 Dublin Core 或是語意標頭皆必須保留對各領域擴充的能力，讓該領域的專家去定義符合其應用需求之特有屬性。本研究則以英語教育領域作為雛形探討的對象。

三、結果

本研究架構延伸文獻探討中 Han 與 Zaine【16】【20】的 VWV 架構，並加以修正而成。主要突破在於：（1）提出一個涵蓋於全球資訊網實際文件一般屬性需求，而且具有足夠彈性，可以滿足全球資訊網上不同領域應用的 Meta Data 定義，作為一個可能為一般超文字文件之提供者所能接受與遵循的上層資料結構。（2）提出虛擬全球 Meta 資料庫的概念，利用現行存在

的搜尋引擎與目錄服務的機制，具體成形一個可行的多層次資料庫的組織架構。(3) 提出在網際網路資料發掘的過程中，可以加入領域與改寫規則與專門字彙涵意的資料庫等智慧型系統的輔助，讓資料發掘的過程更有效率以及彈性。圖一為本研究之架構，圖中層級 1 至層級 2 為各網路資源蒐集與整理者將其資料庫中具有結構化與描述性特質的 Meta Data 網頁描述的資料分享出來，並進行篩選與處理語意的動作，以組成一個虛擬的全球 Meta Data 資源資料庫。本研究對於所謂的 Meta Data，在一般網際網路的通用屬性上，係改進自前述文獻探討中 Desai【6】所建議之 Semantic Header，並融合部份美國 OCLC 與 NCSA 對於擴充超文本語言 HEAD 標記段落中的 META 標記子段落制定之 Dublin Core 十二項基本資料項目定義，再加上針對應用領域屬性擴充的資料項目，並改進涵蓋範圍（Coverage）、辨識字串（Identifier）、相關人員（RespAgent）、與其他文件的相關性（Relation）等定義。

對於圖一中的 Meta Data 概念樹的定義，有別於 Han 對一般資料庫作資料發掘時使用單一屬性概念樹的概念，本研究同時使用單一屬性概念樹及複合屬性多重概念樹來幫助資料發掘的進行。所謂的複合屬性概念樹，是指概念樹表示 Meta Data 某一組屬性（即複合屬性）之屬性值的層級觀念。其原因是由於我們在遵循 Desai（1997）語意標頭的做法，在設計 Meta Data 項目時，為了讓一個 Meta Data 項目能夠同時廣泛表達各種意含，給予其屬性較大的彈性所致[1]。

在雛形實作上，本研究對使用者介面，以動態伺服器網頁（ASP）程式語言（共約五千行），建立全球資訊網環境中以視窗的圖形化介面，供使用者點選並輸入需求，提供其方便的介面操作（如圖二）。資料庫內容，則以網際網路 Meta Data 資源資料庫，以 EXCEL 巨集產生 200 筆虛設網址資料，分別存於十三個關連表格中，使用的資料庫管理系統軟體為 Microsoft 公司的 SQL Server。改寫規則部份，本研究利用國外發展已有一定時日的 JESS（為一個專家系統 Shell）。加強其與全球資訊網資料庫方面的連結，以獨立的 JAVA Applet 嵌入 ASP 中，並在 SQL Server 中建立知識庫處理，以進行對於所有改寫規則的處理。目前在雛形中共存有一般常識型的改寫規則八十四條、一般網際網路的改寫規則十條與應用領域的改寫規則十條。整個系統之雛形環境如圖三。

此系統共有四大模組：在模組一、二中，讓使用者作一般全球資訊網 Meta Data 屬性項目與應用領域的選取，並遞迴找出應用領域之子領域；並以類似 QBE 方式來縮小有興趣的群體範圍。但是因未實作專有名詞辭典，所以使用者之條件內，不可下資料中沒有涵蓋的值。在模組三中，運用專家系統對於網際網路的知識、一般常識與選定應用領域之企業知

[1] 以涵蓋範圍（Coverage）這個複合屬性的 MetaData 項目為例，其實包含涵蓋範圍歸屬領域與涵蓋範圍對應值格式、涵蓋範圍對應值三個屬性。涵蓋範圍歸屬領域的不同便可能同時包含：閱讀年齡限制、適合何種收入者閱讀、地理上的文件有效的涵蓋範圍等不同的資料，因而對應不同的概念樹，

識，做適當的過濾與改寫。在模組四中，建立起始表格，使用者了解所有選取屬性之相關概念樹資訊，考慮其需求，設定門檻值。經過與 Han 類似的概念樹演算法予以抽象化，再對所欲探討之特性或區別規則及有興趣的屬性，設定作為目標或對照組的屬性值。此時屬性值並不一定為概念樹底層樹葉節點，而視其選擇之門檻值，而可能為上幾層之值，若使用者不滿意，而希望選擇較低層次的值，可用「上一步」的方式回到選擇門檻值的地方，放寬門檻值，重新選擇較大數值。最後系統從其中找出知識，並加以解釋。透過本雛形「上一步」「下一步」的使用方式，使用者可以從同一棵概念樹相同概念層次的不同屬性值或屬性抽象值反覆的選擇中，得到多個同一棵概念樹相同概念層次的規則。而不必像 Han 之演算法，即使是同一棵概念樹相同概念層次的相關規則發掘，每一次資料發掘都必需重新從形成起始表格開始作起。

四、討論

本研究之貢獻如下：

(一)學術上的貢獻：

(1)在架構上：(a)對於 Han 的 VWV 架構的來源，明確定義出網際網路可以提供抽象化之上層資料來源，加以改進；(b)提出更完整 Meta Data；（c）全球資訊網 Meta Data 必需使用複合屬性來有彈性並完整之保存相關資料項目資訊，本研究提出網際網路上複合屬性抽象化之問題的處理方法與機制；(d) 提出在資料發掘中加入改寫規則以進行查詢最佳化，加速發展過程。(2)系統實作上：(a) 由文獻所知，Han 與 Zaine 並未真正在全球資訊網環境下建立資料發掘雛形，本研究可能是以概念樹演算法進行全球資訊網的資料發掘雛形的先驅之一；(b) 將精靈「上一步」「下一步」的觀念列入概念樹演算法的實作中，讓使用者一次可對於其前段的變數修正，發掘出較多的知識；(c) 運用 QBE 具有親和性的使用者介面，讓使用者較易下達其需求；(d) 雛型系統四個部份均相當的獨立，彼此的連結均透過資料庫作為資料傳輸的媒介，可以依需要而抽換任一模組；(e) 系統具備擴充性，可以輕易的加入新的知識。

(二)實務應用上的貢獻：

（1）對於目錄服務或搜尋引擎的管理者，本雛型提供一個可以找出合理分配其目錄項目分群分組與瀏覽次序之合理資料來源。(2) 對於一般使用者（如英語教育業）可以瞭解所關心之應用領域在全球資訊網上的相關網頁，是依照何種知識或規則的方式存在。(3) 經由本研究的建議，政府可以找出作為關於全球資訊網的規則與知識的來源。

但本研究雛形實作有以下限制：

（1）由於資料庫選擇 SQL Server 因此 JESS 在資料庫連結上採取較缺乏效率的 JDBC-ODBC Bridge 的連結方式。(2) 尚欠缺完整的由企業規則經由專家系統產生改寫規則的功能。(3) 在某些模組之間，尚無法做到完整的與「下一步」之連結。(4) 沒有實作專業辭典，無法允許使用者在 QBE 中下達專門辭典的條件。(5) 系統內所建立之常識、一般網際網路規則、應用領域規則之知識庫尚未建立完整，

無法應付現實資料發掘之需求。（6）關於 QBE 之功能在雛形中尚有限制。（7）在模組四中，雖有讓使用者可以修改 SQL 起始查詢的文字方塊，但卻欠缺文法的相關編譯器檢查。（8）對於複合三個屬性形成的概念樹，若因 Schema 不同而有不同對應的情況，在雛型中並未實作。

建議後續研究發展方向如下：

（1）可以就如何利用全球分散運算與儲存之方式，建立虛擬全球 Meta Database 作相關探討。（2）可以繼續找尋不同的應用領域，規範其從屬關係與屬性。（3）目前無論本研究或是 Han 演算法，都仍在效率上無法滿足實務的需求，因此仍有待研究更具效率的演算法。（4）增加具有完整專家系統功能的自動改寫企業規則模組。（5）對於複合屬性概念樹更有效的處理機制，以及對於缺值自動參照補值的處理進行相關探討。（6）對於網際網路上多媒體資料的處理進行相關探討。（7）對於將研究對象由網際網路擴充到企業網路作相關的研究與探討。

五、計畫成果自評

研究內容與原計畫甚為相符，由上述結論可以看出達成之預期目標包含：創新模式架構之提出及實驗雛形之建立，並在此過程中培育人才。學術價值高，也有應用價值，可供網路業者參考，相關之成果正在改寫成論文，期待發表於國內外之期刊。

六、參考文獻

【1】 Agrawal, R., Ghosh, S., Imielinski, T.,Iyer, B. and Swami, A., “An Interval Classifier for Database Mining Applications,” *Proceeding of the 18th VLDB Conference,* Vancouver, Canada, August, 1992, pp.560-573.

【2】 Agrawal, R., Imielinski, T. and Swami, A., “Database Mining: A Performance Perspective,” *IEEE Transactions on Knowledge*, Vol.5, No.6, December 1993, pp.914-925.

【3】 Agrawal, R. and Srikant, R., “Mining Sequential Patterns,” *IEEE 11th International Conference on Data Engineering*, Taipei, Taiwan, March, 1995.

【4】 Bowman, M.,Danzig, P. B.,Manber, U. and Schwartz, M., “A Scalable, Customizable Discovery and Access System,” *Technical Report CU-CS-732-94*, Department of CS, University of Colorado, Boulder, July, 1994.

【5】 Chen, M. S., Han, J. and Yu, P. S. “ Data Mining : An Overview from Database Perspective,” *IEEE Transaction Knowledge and Data Engineering*, December, 1996, pp.886-883.

【6】 Desai, B. C., “Supporting Discovery in Virtual Libraries,” *Journal of the American Society for Information Science*, Vol.48, No.3, Mar 1997, pp.190-204.

【7】 Etzioni, O., “The World-Wide Web: Quagmire or Gold Mine?” *Communications of ACM*, Vol.39, No.11, November, 1996, pp.65-68.

【8】 Fayyad, U. M., “Data Mining and

Knowledge Discovery : Making Sense out of Data," *IEEE Expert*, Vol.11, No.5, October, 1996, pp.926-938.

【9】 Frawley, W. J., Paitetsky-Shapiro, G. and Matheus C. J., "Knowledge Discovery in Databases : An Overview, " *Knowledge Discovery in Databases*, California, Edited by G. Paitetsky-Shapiro and W.J. Frawley, AAAI/MIT Express, pp.1-30.

【10】 Grupe, F. H. and Owrang, M. H. "Data Base Mining Discovering New Knowledge and Cooperative Advantage," *Information System Management*, Vol.12, No. 4,Fall, 1995, pp.26-31.

【11】 Han, J., Cai, Y. and Cercone, N., "Knowledge Discovery in Databases : An Attribute-Oriented Approach," *Proceeding of the 18th VLDB Conference*, Canada, August, 1992, pp. 547-549.

【12】 Han, J., Y. Cai, N. Cercone, and Huang, Y." Discovery of Data Evolution Regularities in Large Databases," *Journal of Computer and Software Engineering, 3(1)*,1995,pp.41-69.

【13】 Han, J. and Fu, Y., "Discovery of Multiple-Level Association Rules from Large Databases," *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95),* Zuich, Switzerland, September 1995, pp. 420-431.

【14】 Han, J., Koperski, K. and Adhikary, J., " Spatial Data Mining: and Challenges," *1996 SIGMOD'96 Workshop. on Research Issues on Data Mining and Knowledge Discovery (DMKD'96)*, Montreal, Canada, June 1996.

【15】 Han, J., Nishio, S. and Kawano, H., " Knowledge Discovery in Object-Oriented and Active Databases," *Knowledge Building and Knowledge Sharing* , Edited by F. Fuchi and T. Yokoi, Ohmsha, Ltd. and IOS Press, 1994, pp. 221-230.

【16】 Han, J., Zaine, O. R., and Fu, Y., "Resource and Knowledge Discovery in Global Information Systems: A Scalable Multiple Layered Database Approach," *Proc. of a Forum on Research and Technology Advances in Digital Libraries (ADL'95)*, McLean, Virginia, May 1995.

【17】 Schwartz, M. F., Emtage, A., Kahle,B. and Neuman,B. C., " A Comparison of Internet Resource Discovery Approaches," *Comput. Syst.*, No.5, Fall, 1992, pp. 461-493.

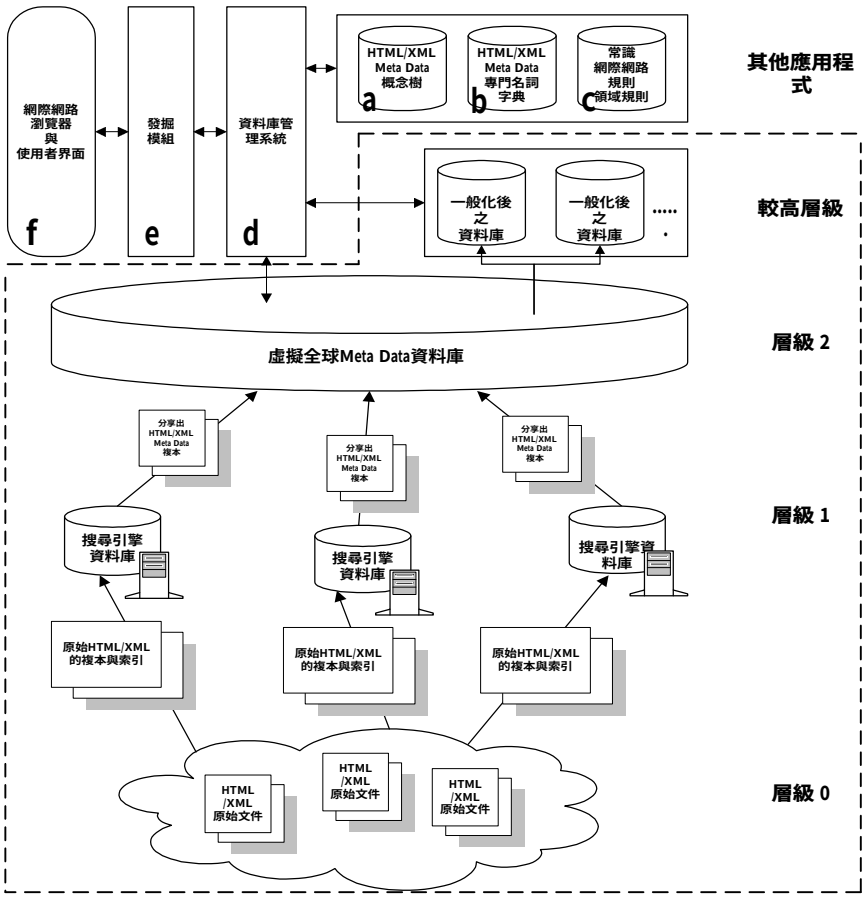
【18】 Weibel, S., Godby J., and Miller E., " OCLC/NCSA Metadata Workshop Report," 1996, Available from *http://www.oclc.org:5046/oclc/research/…rences/metadata/dublin_core_report.html/*.

【19】 Yoon, J. P. and Kerschberg, L., "A Framework for Knowledge Discovery and Evolution in Databases," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No.6, December, 1993, pp.973-979.

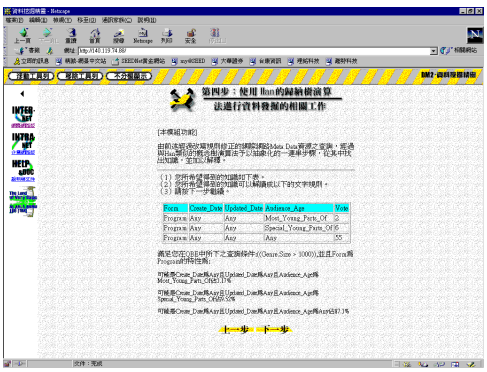
【20】 Zaine, O. R., “Resource and Knowledge Discovery from the Internet and Multimedia Repositories,” Ph.D. Thesis, Simon Fraser University of Computer Science, 1999 .

則：用學生修課的資料作分析，”淡江大學資訊工程研究所碩士論文，民國八十四年。

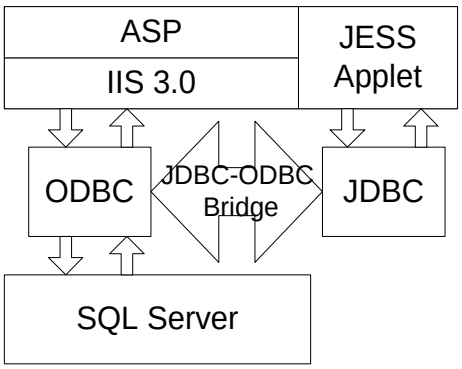
【22】 薛如芳，“以歸納學習法自關聯式資料庫中發掘知識，“ 交通大學資訊工程研究所碩士論文，民國八十四年。



圖一 本研究理論架構



圖二 雛形之 WWW 環境介面圖



圖三 雛形環境圖

【21】 周立平，“從資料庫中發現法